



International Actuarial Association  
Association Actuarielle Internationale

# Artificial Intelligence Governance Framework

AI Task Force  
November 2025

## **IAA Paper**

### **Artificial Intelligence Governance Framework**

#### **A comprehensive governance framework on artificial intelligence for actuarial work**

This paper was prepared by the Artificial Intelligence Task Force (AITF) of the International Actuarial Association (IAA).

The IAA is the worldwide association of professional actuarial associations, with several special interest sections and working groups for individual actuaries. The IAA exists to encourage the development of a global profession, acknowledged as technically competent and professionally reliable, which will ensure that the public interest is served.

The role of the AI Task Force is to deliver on the Statement of Intent for IAA Activities on Artificial Intelligence (SOI) as adopted by Council on 8 March 2024.

The paper was authored by a drafting group appointed by the AI Task Force.

This paper has been approved for IAA publication by the AI Task Force and the Executive Committee in accordance with the IAA's Communications Policy.

This paper is published by the IAA solely to encourage understanding and debate of the issues raised therein. For the avoidance of any doubt, this is not an International Standard of Actuarial Practice (ISAP), nor does it set standards or requirements which any individual or organization is expected to consider or observe, or with which they are expected to comply. This is the case notwithstanding any language in the paper which, but for this clause, might suggest otherwise. This statement takes precedence over any such wording.



International Actuarial Association  
Association Actuarielle Internationale

**Tel:** +1-613-236-0886 **Fax:** +1-613-236-1386

**Email:** [secretariat@actuaries.org](mailto:secretariat@actuaries.org)

605 - 75 Albert St, Ottawa ON K1P 5E7 Canada

[www.actuaries.org](http://www.actuaries.org)

## Table of Contents

1. Introduction.....	1
1.1 Purpose .....	1
1.2 Importance of AI Governance.....	2
2. Key Components of AI Governance Framework .....	3
2.1 Roles and Responsibilities .....	5
2.2 Board of Directors.....	6
2.3 Committees and Policies .....	6
2.4 Key Functions.....	7
2.5 Model Owner .....	7
2.6 Model Risk Ratings.....	8
2.7 Key Governance and Risk Management Processes.....	9
2.8 Independent Validation of an AI Model.....	9
2.9 Applicability of Framework to Third-Party Vendor AI Models and Data .....	11
2.10 Human Supervision and Oversight.....	12
3. Governance Over an AI System or AI Model Lifecycle.....	12
3.1 Overview .....	12
3.2 Designing the AI System .....	13
3.2.1 Bias, Fairness and Discrimination.....	14
3.2.2 Transparency and Explainability.....	15
3.3 Developing the AI System .....	16
3.3.1 Gathering and Preparing the Data.....	17
3.3.2 Training and Evaluating the AI Model .....	19
3.3.3 Documenting the AI System.....	21
3.4 Approving the AI System.....	22
3.5 Implementing the AI System.....	22
3.6 Ongoing Monitoring of the AI System .....	23
4. Additional Considerations.....	25
4.1 Training and Education.....	25
5. Conclusion .....	26
Appendix: Definitions – Bias, Fairness and Discrimination .....	27
References.....	34

## 1. Introduction

### 1.1 Purpose

This paper aims to provide educational material that helps actuaries in safeguarding responsible artificial intelligence (AI), while raising awareness of the risks that need to be managed when designing, developing, implementing and using AI models and AI systems. Actuaries have long been at the forefront of managing uncertainty, utilizing a combination of skills in areas such as probability theory, advanced mathematics, statistics, economics, finance and, importantly, professionalism. As key players in decision-making within the financial industry and the field of social protection, actuaries’ concerns about managing risks appropriately and contributing to societal well-being have become even more relevant with the rise of AI. AI is no longer an emerging topic; it has already found a place within the actuarial profession, particularly in data analytics, predictive modelling and risk management, and its impact on actuarial practice is only bound to grow. Given the distinctive nature of the actuarial profession – encompassing technical skills, ethical standards and professionalism – actuaries are well positioned to contribute to the development of AI systems and to oversee their principal’s overall approach to AI.

This paper will build upon traditional governance areas impacted by AI. The subsequent sections will describe best practices for governing and mitigating risks related to data, modelling and the outcomes produced by AI systems once deployed. The intended readers of this paper are primarily actuaries and professionals involved in actuarial work who are seeking to enhance their understanding of AI governance. In this paper we will use the Organization for Economic Co-operation and Development (OECD) definitions of an AI system and an AI model,<sup>1</sup> highlighted within the following call-out box:

#### Definitions

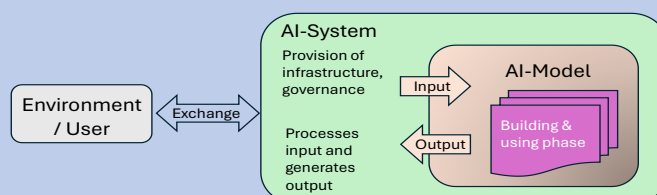
##### AI System

“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.”

##### AI Model

“An AI model is a core component of an AI system used to make inferences from input to produce outputs.”

An AI model is distinguished from a traditional model by having adaptive and autonomous features.



<sup>1</sup> OECD (2024), *Explanatory Memorandum on the Updated OECD Definition of an AI System*, [https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system\\_3c815e51/623da898-en.pdf](https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf)

Note that several of the references included in this paper relate to the use of AI within an insurance environment. However, the content of the paper is not limited to insurance, but rather is meant to be broadly applicable to the use of AI in other environments as well.

## **1.2 Importance of AI Governance**

The importance of developing and adopting a governance framework for a more AI-focused landscape is crucial to actuarial work. The complexity of AI introduces a range of challenges that are considered in this paper. One example is the potential for bias<sup>2</sup>, regarding gender or race. This issue is amplified by the large datasets used to train AI systems, which often incorporate historical, human-influenced data, including language, societal feedback and human interactions. Unintentionally, algorithms can inherit and perpetuate the characteristics of the training data, as is frequently observed in Large Language Models (LLMs). Since models do not “understand” the underlying issues, they are shaped by how data is processed, the design of the model, implementation decisions and output configurations. For instance, the pricing and risk assessment processes actuaries are involved in may inadvertently capture protected characteristics of customers and hence result in discrimination,<sup>3</sup> and privacy issues resulting from the flow of generated output that contain protected information (although conversely, confidence that such discrimination has not occurred requires knowledge of the protected characteristics).

First, gaining a comprehensive understanding of how to manage and mitigate the risks associated with the use of data and modelling practices is increasingly important. Robust governance structures and controls are essential to ensure that the use of data and the implementation of actuarial models are conducted appropriately. Having an AI governance framework in place, and in conjunction with existing organizational and IT system governance edicts, will help address potential errors, identify weaknesses, maintain consistency across actuarial services and ultimately support better decision-making. Consistently applying this governance framework throughout an organization will foster productive dialogue among relevant stakeholders, providing high-quality information, models and assumptions that are fit for purpose.

Second, an AI governance framework can help actuaries meet regulatory requirements specific to their entity’s region and sector. Increasingly, regulators around the world are focusing on the risks associated with AI system implementation, particularly in areas such as data governance, model governance, testing and validation. By acting as facilitators of proper governance framework implementation within their organizations, actuaries can further contribute to mitigating existing risks, enhancing productivity and improving organizational efficiency. This, in turn, allows for a stronger focus on delivering high-value analysis aligned with business needs.

This paper considers those documents that the IAA has already developed on the subject of governance of actuarial work and, in particular regarding governance of models, ISAP 1

---

<sup>2</sup> See Appendix for discussion on Bias, Fairness and Discrimination.

<sup>3</sup> See Appendix for discussion on bias, fairness and discrimination.

International Standard of Actuarial Practice 1 *General Actuarial Practice* (2018).<sup>4</sup> Parts of ISAP 1 that are particularly relevant to this paper are:

- Section 2.1 Acceptance of Assignment;
- Section 2.2 Knowledge of Relevant Circumstances;
- Section 2.3 Reliance on Others;
- Section 2.5 Data Quality;
- Sections 2.6 to 2.9 on assumptions and methodology;
- Section 2.10 Model Governance;
- Section 2.11 Process Management; and
- Section 3.1 Communication – General Principles.

Note that ISAPs are model standards of actuarial practice developed by the IAA, and it is up to relevant standard setters to incorporate the content of ISAPs into their own standards as they deem appropriate. As such, the requirements of the ISAPs are not directly binding on any actuary as they are only binding to the extent adopted by a relevant standard setter. A standard of actuarial practice is a statement of behaviour expected of actuaries operating within a specified context. However, this paper is purely educational and no statement in it requires compliance. Nonetheless, the relationship between this paper and ISAP 1 is relevant, and this paper complements the guidance provided in ISAP 1 about governance of AI models by providing further considerations on the subject the actuary may want to observe. This framework should be read as an actuarial-specific educational supplement to existing international AI governance document (e.g., ISO 42001, NIST AI RMF, OECD AI Principles). It is not intended to duplicate those documents, but to highlight areas where actuarial work might require additional or different governance considerations.

Significant AI governance issues relate to what is meant by the terms “bias”, “fairness” and “discrimination”. The Appendix provides definitions of these in this context, as well as considering associated issues.

The depth and frequency of governance activities should be proportionate to the risk and potential impact of the AI system, ensuring resources are focused where risks to customers, the public or the business are greatest.

## 2. Key Components of AI Governance Framework

To understand the key components of an AI governance framework, actuaries may first choose to reference relevant law or existing regulatory guidance for model governance or model risk management (MRM) as background information. For example, the U.S. Board of Governors of the Federal Reserve System’s *Supervisory Guidance on Model Risk Management*

---

<sup>4</sup> [https://actuaries.org/actuarial\\_practices/isap-1-general-actuarial-practice/](https://actuaries.org/actuarial_practices/isap-1-general-actuarial-practice/). Note that at the time of producing this paper the IAA is considering amendments to ISAP 1, and parts of ISAP 1 relevant to the subject of this paper may change in the near future. The reader is encouraged to confirm whether a new ISAP 1 is issued by the IAA and to what extent the changes to ISAP 1 have an impact on the content of this paper.

(MRM), promulgated in 2011, provides such a reference. This guidance established expectations for a bank's adoption of MRM practices in the U.S. Since 2011, many financial services firms throughout the world, including insurance companies, have developed MRM frameworks that have adopted or adapted such guidance to align with the size and complexity of their organizations.

Furthermore, the actuary may also reference relevant binding actuarial standards of practice, among them ones issued by their standard-setter that may be substantially consistent with ISAP 1 or address similar considerations.

Foundational to a model governance framework is the definition of a model. For the purpose of this paper, we will use the IAA's definition of a model.<sup>5</sup> In some cases, model definitions also include references to the components of a model that need to be addressed within a model governance framework, namely an input component (e.g., data and assumptions), a processing component (e.g., a calculation engine or algorithm) and an output component (e.g., a report or data and information relied upon by users of the output).

It is clear from the IAA's definition of a model that AI models would be captured within this definition and thus be subject to the expectations set out in established model governance frameworks. While AI models pose different challenges from a model governance perspective than more traditional models, the overall principles and components of a model governance framework apply.

"Model governance", in its simplest form, refers to the roles and responsibilities of individuals involved in the design, development, implementation and use of models, including the senior leaders responsible for the oversight of model risk within an organization and the policies that guide them. Such roles and responsibilities typically extend to those charged with the effective challenge of models within an organization, such as within a "three lines of defence" structure, including leaders from the business, risk management and internal audit functions, and the organization's board of directors.

However, the term "model governance framework" often extends beyond this simplest form of governance. Such frameworks include both the establishment of management controls over a model's development, implementation, ongoing use and eventual retirement, and the risk management processes that serve to mitigate model risk throughout a model's lifecycle, capture an organization's model risk appetite and encourage internal audits that seek to confirm that the organization's management of model risk is performed in a manner consistent with expectations.

Organizations that have established model governance frameworks (hereinafter used interchangeably with the term "model risk management frameworks") have typically introduced policies or guidelines for the purpose of clearly articulating MRM expectations, including:

---

<sup>5</sup> "A representation of some aspect of the world. The model produces a set of outputs from inputs in the form of data and other information, assumptions, and parameters. Inputs and outputs may be qualitative or quantitative. The model is functionally designed by a specification that describes the matters that should be represented, the inputs and the relationships between the inputs, and the resulting outputs. The model is technically designed and implemented through a set of mathematical formulae and algorithms."

- The roles and responsibilities associated with model governance, including escalation protocols;
- Key terms used within the framework and their definitions;
- The organization’s model risk rating methodology for the purpose of identifying “high risk” models;
- Key governance and risk management processes;
- The applicability of the framework to third-party models; and
- Expectations for model documentation and model risk reporting.

When developing, managing the risk of or using AI systems, actuaries are encouraged to consider the applicability of any internal policies or practices addressing model and data governance, and also applicable standards of actuarial practice that are substantially consistent with ISAP 1 Section 2.10. In some cases, actuaries will observe that the AI system they are currently developing or using or planning to develop or use does not fit the definition of a model and thus would not be swept into an existing model governance framework. In this circumstance, governance guidelines may separately exist as a reference, or the model governance framework could be consulted for guidance on good governance practices. In addition, actuaries may turn to existing data governance frameworks, often adopted in accordance with regulations established for the safe and secure use of data.

The following sections highlight the key components of a model governance framework while highlighting actuarial considerations when designing, developing, implementing or using AI.

## **2.1 Roles and Responsibilities**

As stated, a strong model governance framework begins with the clear articulation of roles and responsibilities of the senior leaders involved in the development or appropriate use of models, including the management of model risk. Such an articulation of roles might include those who are responsible for strategy/oversight and management on the enterprise level, such as the board of directors, senior management committees, the Chief Risk Officer (CRO) and internal audit on the one hand, and on the other hand those who are responsible for operational-level governance bodies such as business functional leaders responsible for the use of models (including the head of IT and data/AI scientists), model owners who have the responsibility for the appropriate development and use of the model throughout its lifecycle.

Typically, a strong model governance program will reflect a risk-based view, with model risk oversight following a three-level approach:

1st Level	Strategy/Oversight	Board of directors <sup>6</sup>
2nd Level	Management	Committees and Policies, Executive Management, Key Functions
3rd Level	Operational	IT, end-user training and support management

---

<sup>6</sup> “Board of directors” refers to the highest body within the entity that is charged with governance of the entity.

The rapid evolution and the usage of AI systems may require an AI-related adjustment of the existing governance structure, depending on the new material risks arising as a result of the use of AI systems within an organization. For example, the multi-disciplinary nature or the use of third-party components of AI system development, and the dynamic nature of the AI systems themselves, pose increased accountability risk whereby it may not be clear who is responsible for the outcomes of the AI system, and at what stage of the model lifecycle these responsibilities are held. In addition, the governance framework should establish clear channels for communicating AI-related matters to key external stakeholders, such as regulators and external auditors, whenever necessary.

As a result, new roles and responsibilities may need to be established to effectively manage these additional risks and the widespread use of AI throughout the organization.

Actuaries currently hold a variety of leadership roles with responsibility for the development and appropriate use of high-risk models, including AI systems. In some cases, actuaries serve as members of teams reporting to such leaders. As a result, the actuary involved in the development or use of AI is expected to understand the organization's expectations for the management and mitigation of model risk at all levels within the organization.

## **2.2 Board of Directors**

The board of directors also bears the ultimate responsibility for the use of AI and thus for an appropriate governance framework that includes developing a new framework and/or updating existing ones (e.g., IT; risk management, including MRM) with clear communication about the strategy for and policies relating to the use of AI within the organization. To accomplish this, the board of directors should have sufficient knowledge of how AI is being used and its potential risks. For this reason, the board should receive adequate and continuous training on AI risks and opportunities and consult with available and appropriate subject matter experts. In addition, relevant policy documents approved by the board provide the framework for the development and the usage of AI in the organization (e.g., goals, principles, values, processes, requirements, responsibilities). The risk management framework approved by the board should address all material AI activities and their related risks, and in turn ensure that the use of AI systems aligns with the organisation's vision and values. The board should receive regular updates through an emerging risk management process so it can maintain a continuous, holistic view of AI use cases and their potential risks.

## **2.3 Committees and Policies**

For existing senior management committees consisting of key governance roles (e.g., CRO, CFO, functional leaders), organizations may need to revise current policies and procedures to ensure the development and use of responsible and trustworthy AI. Especially for high-impact AI, an entity may evaluate whether it is worthwhile to implement a standalone AI/Data/Ethics Committee, which takes care of the development, deployment or procurement of AI systems, or whether these responsibilities are best managed by existing committees. Regardless of the decision to create new or utilize existing committees for the oversight of AI, an organization

should consider including a minimum number of members with a “fit and proper”<sup>7</sup> profile on AI. It is important that those accountable for the AI system’s deployment understand enough about how the system works to be able to identify, and mitigate, the associated risks.

## **2.4 Key Functions**

For certain key oversight functions (i.e., risk management, actuarial, compliance and internal audit functions within an insurance organization), AI-related issues need to be integrated in the corresponding positions in an adequate manner respecting the very different roles and responsibilities of each function. Given a high-risk AI system, the impact on the risk management (RM) function (including MRM) could be very different compared to the impact on the compliance function. While the RM function has to find a measurement for the risk, which could be very time-consuming and complicated, the compliance function needs to ensure that the deployment and the operation of the AI system is in line with existing internal policies and external regulations. In principle the need for the adjustment for each function can be evaluated using an appropriate gap analysis.

It is important to note that some organizations are sufficiently small to not have formal MRM practices in place. In these instances, internal audit may adopt certain “second line” functions to evaluate the effectiveness of governance and controls over high-risk models, and implicitly AI systems, in accordance with the organization’s policies and guidelines.

## **2.5 Model Owner**

The model owner owns all aspects of the model, including development, performance monitoring, compliance and maintenance. The model owner ensures the model meets specific business objectives, aligns with strategic goals and business objectives, complies with regulations and internal policies, and remains effective in its purpose. In addition, it provides value throughout the model’s entire lifecycle, from inception through decommissioning. The model owner remains accountable for the performance, compliance and ethical operation of the end-to-end AI system in which the model is embedded. The model owner is responsible for the data within the models they develop or use and should approve how that data is defined and utilized. Usually, the model owner is responsible for the complete value chain (cf. the illustrative example below).

Responsibilities of the model owner could include:

- Defining the responsibilities for the model stakeholders, including data scientists, IT, compliance and business units;
- Assigning stewards responsible for maintaining, updating and monitoring models throughout their lifecycle;
- Establishing and/or approving workflows for key stages of model development, such as pre-deployment, post-deployment and significant updates;

---

<sup>7</sup> “Fit” means that the professional qualifications, knowledge and experience of the individuals and are adequate to enable sound and prudent management. “Proper” means the individuals and are of good repute and integrity.

- Setting up accountability structures to ensure that the models are used responsibly and in alignment with organizational values;
- Assembling an appropriate cross-functional team (data scientists, compliance officers, IT security and business leaders) to support governance, and working closely with the team to ensure the model adheres to internal policies and external regulations;
- Regularly updating stakeholders on model performance, risks and any significant changes;
- Deciding when the model should be retrained, updated or retired (based on organizational requirements including performance, cost and ongoing relevance or per contractual obligations specified with third-party agreements for data usage or model/AI usage);
- Managing models and approving their definitions and requirements;
- Reviewing and approving specifications related to the model standard and revisions to ensure that any key changes are sufficiently understood and their impact is fully assessed and integrated; and
- Ensuring consistency and quality across different types of models and data (e.g., actuarial and other assumptions).

The model owner's responsibility includes close alignment with the IT- Security Officer – and, where applicable, the broader IT infrastructure team – who are responsible for safeguarding digital assets, networks and data from cybersecurity threats. Furthermore, the model owner should collaborate with IT to ensure that data and model security requirements are embedded within the entity's information security management framework, covering both preventive and responsive measures. Model security should encompass the integrity and confidentiality of inputs, outputs and intermediate data, as well as the complete processing workflow. This includes securing storage environments, access controls, audit logging and monitoring for security incidents throughout the Machine Learning Operations (Lops) lifecycle.

## **2.6 Model Risk Ratings**

An important tool within a strong model governance framework is the model risk rating methodology. When appropriately developed and applied, this tool enables the actuary to determine whether the AI system that they are developing or using, initially determined to be within the scope of the model governance framework, is categorized as high-, medium- or low-risk. While risk rating methodologies vary by organizations, they typically include risk criteria such as adverse financial impact and degree of complexity, among other characteristics to assess risk. For example, typical actuarial risk criteria include the potential to impact reserving, solvency and capital, and the potential to impact pricing, underwriting, claims processes and fair treatment of beneficiaries. Model governance frameworks often have more significant governance requirements for high-risk models, with low-risk models typically requiring minimal independent oversight.

Since AI systems pose different risks to organizations and their customers than traditional models, model risk rating methodologies likely have or need to evolve to adapt to the changing model risk environment. For example, such methodologies might consider the degree of

transparency and explainability, the level of autonomy, and the degree of reliance on third-party data or systems.

## 2.7 Key Governance and Risk Management Processes

The MRM processes captured within a model governance framework typically include:

- Model risk identification processes via the application of the risk rating methodology, which are frequently performed in cooperation with functional leaders and model owners;
- Model risk assessments via independent model validations and ongoing model performance monitoring;
- Risk mitigation via the resolution of validation outcomes (if any) and the establishment of additional model control processes; and
- Risk reporting to business leaders and the board of directors.

Actuaries that are developing AI systems or AI models, or performing independent validation of these systems, may need to focus more heavily on input data and outcomes testing than when developing or independently validating traditional actuarial models. The dynamic nature of AI systems is expected to mean that these processes may need to be reviewed, or validated, more regularly post-deployment than traditional actuarial models. See Section 3 for a discussion of several unique considerations when developing or validating AI systems, including addressing data quality and privacy, the transparency and explainability of both data and algorithms, and outcomes testing.

## 2.8 Independent Validation of an AI Model

When it comes to model validation, actuaries are encouraged to consider ISAP 1 Section 2.10.2 on model validation. Validation in the context of this governance framework is intended to provide documented evidence that a model shows certain desired properties (e.g., generalization, accuracy) and does not show undesirable properties (e.g., bias, overfitting). This validation should be performed by individuals who are suitably knowledgeable and trained to perform the assessment but did not develop the model, unless this imposes a burden that is disproportionate to the model risk. The usual evaluation criteria such as performance, stability and reproducibility are consistent across many models, but validation of AI may require an actuary to adapt validation methods to the specific type of AI. These adaptations may be required for the usage of unstructured data, previously manual steps that are performed automatically, the degree of explainability, dynamic updating of the model or missing transparency. It should be noted that this can be specifically challenging with respect to Generative AI (GenAI) models. Some further validation aspects are listed below.

Aspect	Explanation
Goals and requirements	<p>Validating AI models may rely upon objectives that are similar to those for general model validation, including:</p> <ul style="list-style-type: none"> <li>• Model behaviour – e.g., reliability, ethical aspects, user interaction;</li> </ul>

Aspect	Explanation
	<ul style="list-style-type: none"> <li>• Model strengths – e.g., robustness;</li> <li>• Model weaknesses – e.g., biases in the input data; and</li> <li>• Model assumptions – e.g., consistencies.</li> </ul>
Validation quality	Provide documented evidence that a model shows certain desired properties (e.g., accuracy and adequacy, completeness and consistency) and does not show undesirable properties (e.g., inappropriate bias, overfitting, hallucination).
Validation method	<p>Tools and methods required to validate AI models depend on the characteristics of the AI model. Some are:</p> <ul style="list-style-type: none"> <li>• Trained vs. pre-trained models (because of huge data usage);</li> <li>• Observability of stability in training and inference;</li> <li>• Type of unpredictability;</li> <li>• Frequency of model modification and/or recalibration; and</li> <li>• Severity and consequences of model results.</li> </ul>
Model validation checklist	<p>Before starting the validation, the scope should be clear, which could be defined via a checklist. This includes, for example:</p> <ul style="list-style-type: none"> <li>• Which acceptance criteria should be applied?</li> <li>• Does the model fit these acceptance criteria?</li> <li>• How is undesirable model behaviour treated?</li> <li>• How is uncertainty quantified in the model output?</li> <li>• How is known bias in the training data handled?</li> <li>• Are stress, sensitivity or scenario tests appropriate?</li> <li>• Which burden (workforce) is acceptable?</li> <li>• What is necessary for ongoing validation activities?</li> <li>• What training of users is required to ensure proper interpretation of the model results?</li> </ul>
Limitations of validation	Ultimately, the quality of validation is measured by how “well” the adequacy of the models can be assessed. With the help of the validation methods used, it should be possible to assess whether the model fulfills the purpose for which it is to be used. In accepting the validation results, inherent uncertainty levels about the quality of the model should be tolerated.

An important objective of validation is to ensure that the model generalizes well to unseen data upon its deployment in the real world. For this purpose, multiple experiments should be carried out on the model using all three data sets – training data, testing data and, in many cases, validation data – to bring out its best abilities and minimize the changes it undergoes post-deployment. If the deployed AI model is capable of adapting and developing based on live data, these experiments should continue post-deployment to identify when the capabilities and performance of the AI model change over time.

## **2.9 Applicability of Framework to Third-Party Vendor AI Models and Data**

When it comes to reliance on third-party or vendor models, actuaries are encouraged to consider ISAP 1 Section 2.3 Reliance on Others. Third-party or vendor models do not need to follow this exact framework, but it is the actuary's responsibility to confirm that the governance standards of the vendor are sufficient for the purpose of the model. Items in this paper should be considered when using a third-party or vendor model, and additional information may need to be requested to ensure that the model fulfills, and continues to fulfill, the standards for actuarial use. Necessary due diligence should be conducted by the relevant stakeholders, and governance needs to be defined in the contracts. When checking third-party data for any bias, the actuary is encouraged to apply their association's relevant standards that are similar to or substantially consistent with that of ISAP 1 Section 2.3 Reliance on Others. Data privacy, accuracy and regulatory compliance should be adhered to, and ownership clarity should be clearly defined. For external models, the actuary is encouraged to independently validate the model results for accuracy, interpretability, fairness and alignment with actuarial standards. Models need to be auditable and, as such, contractual terms should be established to encourage vendor accountability.

In fact, various GenAI uses involve third-party vendor models, and the supply of such models may be heavily concentrated, which might lead to an elevated level of operational risk that needs to be mitigated in a risk-based and proportional manner. When using third-party data or vendor models, actuaries are encouraged to ensure that the governance standards employed by the vendor are adequate for the intended purpose. While vendors may not follow the same frameworks actuaries use, it is the actuary's responsibility to confirm that the model meets regulatory and actuarial standards, including both professional and practice standards. This involves thoroughly assessing the vendor's governance processes, validating the model's assumptions, and ensuring data quality and transparency. Additional documentation or assurances may be required to confirm the model aligns with actuarial principles and is suitable for deployment.

Ongoing oversight is essential to ensure the model continues to perform effectively and professionally. Processes should be established for monitoring model outputs, identifying and addressing issues, and requesting updates or recalibrations as conditions evolve. Collaboration with vendors is critical for securing necessary documentation, testing outputs for fairness and accuracy, and negotiating any required modifications to align the model with actuarial use cases.

Ultimately, actuaries remain accountable for the use of vendor models and are encouraged to apply professional judgment to evaluate their adequacy. By adhering to a proportional governance approach – tailored to the risk and impact of the model's application – actuaries

can ensure reliable and fair outcomes while maintaining compliance with actuarial standards and regulatory requirements.

## **2.10 Human Supervision and Oversight**

Establishing appropriate human oversight of AI systems is critical for ensuring they function as intended and do not produce adverse effects. Human oversight involves direct involvement in the design, operation, maintenance, adaptation or application of AI systems (e.g., human-in-the-loop). While AI increasingly automates tasks and processes, human involvement remains necessary at all stages of the AI system lifecycle to provide checks and balances.

AI outcomes may be discriminatory; for example, due to the nature of the training data. Consequently, a well-calibrated balance between human judgment and AI automation is essential for certain tasks to mitigate risks and improve outcomes.

The design of human oversight should be proportionate to the specific AI use case, reflecting the nature, scope and complexity of the associated risks, while considering existing governance frameworks. The method of oversight depends on the stage of the AI application lifecycle. For example, during the design phase, developers should be aware of the possibility of biases in training datasets and determine the appropriate level of automation for use cases, such as pricing or underwriting, where human review may be required. Once operational, oversight shifts to monitoring daily processes, controlling system performance, addressing incidents and adapting to changes as per established procedures. This may involve employees reviewing key metrics to assess the AI system's impact on vulnerable groups or requiring human validation of AI outputs before implementation.

## **3. Governance Over an AI System or AI Model Lifecycle**

### **3.1 Overview**

Strong model governance is not achieved by a series of procedures that are applied after an AI system has been developed, but rather as integrated practices and procedures that are applied within every phase of a model's lifecycle, including the design, development, implementation, ongoing monitoring and ultimately retirement of the AI system. Such practices and procedures are designed to achieve a common set of objectives, including fairness and non-discrimination, safety and security, robustness and compliance with regulations. Actuaries have an essential role to play in establishing and maintaining strong model governance over AI models.

For the purpose of establishing and maintaining strong model governance, actuaries are encouraged to be aware of the fundamental characteristics of three broad categories of AI models:

1. Models in which the calculation of the output is transparent and easily explained to the public, such as statistical models (e.g., GLM [Generalised Linear Model], GAM [Generalised Additive Model], Decision Tree);
2. Models in which the relative importance of input characteristics can be explained, but the exact calculation of an individual output is opaque and not easily explained to the public, such as Gradient Boosting Machine [GBM], Neural Networks and Support Vector Machines; and

3. Models in which the relative importance of input characteristics and the calculation of an output is not easily explained to the public, such as BART [Bidirectional and Auto-Regressive Transformers], LLMs and image recognition.

In reality, an AI model may exist on a spectrum between these categories. For example, a decision tree with a single split is considered quite transparent, while ensuring the explainability of a decision tree with thousands of splits normally needs extra effort due to its complexity. Another example are AI models that are used to analyze very large data. The input is well defined, but the output is difficult to explain.

As a result, the type of model chosen will have a significant impact on how readily the organization's objective of transparency and explainability may be achieved, and the types of testing, validation and ongoing monitoring procedures that need to be applied. Actuaries will need to be focused on the unique governance considerations of the types of AI systems chosen during the design and development phases and throughout the applicable lifecycle.

Model governance is particularly necessary to ensure the model is able to appropriately adapt to a changing environment. When a model is deployed in the real world, the data fed to it becomes very dynamic. Apart from the data, there might be changes in technology or business goals, or a drastic real-world event like a pandemic that has an impact on the AI model's performance. Frequent refreshments of the AI model are necessary to keep up with changes in data, technology and regulatory expectations. With many regulators considering implications of AI adoption, monitoring policy developments and maintaining dialogue with supervisors is important. When the AI system is no longer fit for purpose, model governance ensures that the model will be appropriately retired.

### **3.2 Designing the AI System**

Designing an AI system and the associated processes is a critical step to ensure the effectiveness of a strong governance framework. A well-thought-out design begins with determining the alignment between the AI system's capabilities with business needs. When developing specific AI systems or models, effective collaboration and cooperation between teams within an organization are essential. Actuaries, who play a central role in governing the applicability of models, are encouraged to engage with development teams from the outset. This early collaboration supports addressing key considerations, such as identifying limitations and clearly defining intended outcomes throughout the model lifecycle.

Furthermore, using AI responsibly involves navigating a range of complex challenges. For instance, bias in AI systems may stem from factors such as ethnicity or gender. Since AI is trained on extensive datasets influenced by human behaviour, language and societal norms, algorithms can inadvertently inherit and perpetuate the biases present in their training data, as observed in LLMs. These models lack an understanding of underlying issues and rely heavily on predictive methodologies shaped by their design, inputs and outputs. As a result, the rights of individuals – such as women or other marginalized groups – can be disproportionately affected by the way these algorithms operate.

Actuaries responsible for risk assessment models – for example, pricing or reserving models – are encouraged to consider whether protected characteristics are addressed adequately. This proactive approach will lead to a more ethical and effective application of AI systems.

Whether developing or using an AI model, actuaries are encouraged to understand and document the model; the conditions under which it is appropriate for the model to be used, including any limitations of the model for the intended use; the context in which the model will be used; how model inputs will be provided; and how the actuary expects the results of the model will be used (see ISAP 1 Sections 2.10.3 and 2.10.4).

The following sections will explore fairness and transparency in greater detail.

### 3.2.1 Bias, Fairness and Discrimination

Algorithms, in conjunction with their underlying data, can exhibit bias and may result in discrimination (direct or indirect). While the definition of discrimination is highly contextual and varies on a case-by-case basis, institutions and regulators worldwide have attempted to standardize such definitions. For instance, the OECD has recently published recommendations to clarify definitions of AI systems,<sup>8</sup> focusing on responsible stewardship and international cooperation for trustworthy AI. Similarly, the European Commission has developed the Ethics Guidelines for Trustworthy AI,<sup>9</sup> and the Monetary Authority of Singapore previously issued principles aimed at promoting fairness, ethics, accountability and transparency (FEAT).<sup>10</sup>

Other related policies and laws that may positively impact the functioning of AI systems include regulations on data protection (GDPR in Europe), insurance distribution (IDD in Europe) and commercial practices (UCPD, still to be verified).

Direct discrimination involves unfavourable treatment of individuals generally by misusing their protected characteristics as a factor. Indirect discrimination, however, is more complex. Further detail is provided in the Appendix.

Fairness is a less straightforward concept as it depends to a great extent on people's expectations, and hence represents typical practice in the relevant market or society as well as the details of the financial and other products in question; these ideas are developed in the Appendix.

AI systems are subject to feedback loops and dynamic data collection, which can lead to alterations in model behaviour over time. If models influence the underlying data, an algorithm initially considered free of discrimination can eventually evolve into one that is discriminatory.<sup>11</sup>

---

<sup>8</sup> OECD (2024), Recommendation of the Council on OECD Legal Instruments Artificial Intelligence, <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>

<sup>9</sup> European Commission (2019), Ethics guidelines for trustworthy AI,

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

<sup>10</sup> Monitoring Authority of Singapore (2018), *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector*, <https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/feat>

<sup>11</sup> European Union Agency for Fundamental Rights (2022), *Bias in Algorithms: Artificial Intelligence and Discrimination*,

[https://fra.europa.eu/sites/default/files/fra\\_uploads/fra-2022-bias-in-algorithms\\_en.pdf](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf)

Some examples of vulnerabilities to consider when addressing discrimination and fairness, or factors that affect vulnerable groups, include the following:

- **Associated Characteristics:** Age group, low income or poverty, low level of education, migrants, race, gender, etc.;
- **Lifestyle Factors:** Unemployment or homelessness, divorced or single parents, over-indebted persons, prison inmates, accident victims, victims of domestic violence, etc.;
- **Health-Related Factors:** Disabilities, hereditary diseases (e.g., genetically determined conditions), individuals with mental illnesses, etc.; and
- **Digital Illiteracy:** Lack of digital skills, difficulties accessing online or digital services, or challenges understanding services provided online, etc.

In dealing with such factors and related characteristics, individuals may encounter numerous cumulative obstacles, including obstacles to financial inclusion. For example, actuaries may need to implement appropriate preventive measures within their modelling frameworks.

Effective and transparent data governance is crucial for ensuring fair and non-discriminatory treatment of consumers. This involves key elements such as transparency, explainability and robust data management practices. Additionally, principles like robustness and performance, as well as human oversight, are essential for fostering a responsible and meaningful approach to AI, helping to address deficiencies within the data. Algorithms should be designed and refined to prioritize explainability and transparency to uphold fairness, while steering clear of practices that might unduly influence consumer behaviour.

A significant challenge in promoting fairness and non-discrimination in AI systems is to improve outcomes by removing or minimizing any bias and discrimination discovered at any time in the governance process. The key for this is to analyze, evaluate and understand the impact of implemented corrective measures.

### **3.2.2 Transparency and Explainability**

The professional use of AI comes with a set of responsibilities, including transparency, explainability, accountability (to be discussed in Section 3.5 Implementing the AI System) and robustness. Ensuring that actuarial professional ethics align with these responsibilities is crucial, and adopting sound system-building governance is fundamental for explainable and transparent AI. Actuaries need to deal with multiple stakeholders; therefore, ensuring these concepts are addressed accordingly will result in better decision-making and well-understood processes.

Transparency issues arise with opaque systems that are difficult to validate, creating consequences for all stakeholders and customers involved. It is crucial that the way a system works and in what way it is documented are clearly communicated to all parties involved. Disclosing the right amount of information will ensure that the system is well understood, uses the right data sources, and is correctly implemented and thoroughly tested, and its outputs are clearly explained and understandable. In some regulatory environments, such as under Europe's AI Act, GDPR or Solvency II, traditional actuarial models are already somewhat aligned with transparency requirements. Generally speaking, actuarial models have often struggled with being fully disclosed or well presented to a wider audience. This issue stems partly from the complexity of such models and the variety of assumptions needed based on

business requirements. These challenges can be amplified when adopting AI models, where the degree of randomness and uncertainty inherent in AI system operations adds another layer to the methodology. Therefore, it is important to clearly outline the intended purpose of the model, and highlight the accuracy of the algorithms used, the robustness of the infrastructure on which the model operates, the appropriate software testing methods and the necessary documentation that reflects human intervention, choice of assumptions and the degree of uncertainty (consider feature selection, model drift, etc.). Ultimately, one should be able to replicate the results of a fully transparent model. (For more details, refer to the IAA papers on testing and documentation of AI models or systems.)

Explainability in AI often pertains to how model choices and assumptions manifest in the system's reasoning process and contribute to achieving the intended output. The degree of explanation required varies by application. For example, some AI chatbots that do not significantly impact decision-making and pose no consequences to individuals may not require extensive disclosure; their outputs are generally self-explanatory. However, more complex models necessitate robust explainability measures that align closely with the system's intended use and the methodology for deriving its outputs. Numerous explainability methods exist, though they are often complex to implement and require a deep academic understanding. Actuaries are uniquely positioned to serve as a bridge in this context, possessing the expertise to comprehend complex mathematical methods and the dynamics of their specific business cases. Many current methods utilize visualization techniques and can be categorized into two types: local and global metrics. Local metrics, such as Individual Conditional Expectation (ICE), Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP), often focus on the influence of individual features on predictions. Global metrics, on the other hand, assess how features affect the system overall and include tools like Partial Dependence Plots, Feature Importance Stability or Fairness metrics.

Conceptuality, explainability and transparency are integral to fairness and non-discrimination in AI systems, serving as key prerequisites for identifying issues within these areas. These elements foster trust and good governance, particularly in financial services, making explainability a fundamental component for responsible AI use.

### **3.3 Developing the AI System**

Actuaries who are part of an AI system development team may be involved in some or all of the AI system's development steps, from its design to meet the organization's business objectives, to training, evaluating, and documenting the system and/or model. Developers should be aware of the unique model governance objectives of each step of the AI development process.

In some organizations, actuaries may be part of AI development teams that have adopted formalized Development Operations (DevOps) methodologies or are using DevOps systems. Such methodologies and systems focus on both software development (Dev) and technical operations (Ops), and typically guide or support developers with data preparation, training and testing, automated deployment and ongoing monitoring. In such cases, the organization's model governance policies and practices should be consistently applied within the DevOps environment.

### **3.3.1 Gathering and Preparing the Data**

As all models are only as accurate as the data behind the model, it becomes crucial to identify appropriate data (sources, types and formats) to ensure model accuracy and relevance. Data governance applies to all phases of the data cycle, from data capture and preprocessing to data enrichment. It is the responsibility of the AI system's model owner and development team to ensure that it has access to data that is relevant for the AI system's intended purpose and that there are no regulatory or ethical constraints with respect to the collection and/or use of the data. Many countries have data security and privacy regulations, and laws to protect consumers' personal information from misuse and exploitation. When using AI, care and attention should be taken regarding the legal mandates on storage and usage of consumer data that apply to any type of actuarial work.

Similar to all aspects of actuarial work, when actuaries are gathering and preparing data for the development of an AI system or using an existing AI system for their purpose, it is important that they understand the data governance practices of their organization, namely the policies, processes and practices that have already been developed to guide how data is managed, used and protected. The goal is to ensure data quality, security, privacy, and overall compliance with policies and procedures.

Key components of data governance such as those outlined below also apply to work involving AI systems:

- **Policies and Standards:** Guidelines defining data management rules, including access, sharing and retention;
- **Ownership and Stewardship:** The assignment of roles (e.g., data owners, stewards) to ensure accountability and proper data handling;
- **Quality Management:** Ensuring accuracy, completeness, consistency and timeliness of data;
- **Compliance Management:** Alignment with existing local regulations, or regulations that apply to the region in scope (e.g., Europe's GDPR, the California Consumer Privacy Act);
- **Technology and Tools:** Platforms for data cataloging, monitoring and compliance reporting; and
- **Risk Management:** Identifying and mitigating risks, such as data breaches or non-compliance penalties.

As AI models are more of a black box compared to other actuarial models, extra care should be taken to check that the data and its usage do not lead to biased outcomes. This is even more critical for AI models that impact consumers directly, such as in underwriting, pricing and claims settlement. For instance, extra care should be taken to check that the data used in the model is:

- Current and up to date;
- Appropriate, given the intended use of the model;
- Consistent and not contradictory (e.g., data field definitions have not changed over time);
- Reliable (e.g., the origin should be known or traceable);

- Complete (e.g., records and/or attributes are not missing for critical components, or the impact of any missing data is understood and accounted for); and
- Representative of the various constituency groups expected to use or be affected by the output of the model.

The model owner or members of a development team should be transparent and clear with respect to the collection of the data, and should check that the data is secure (tamper-free), collection of the data has been authorized (requisite permission has been obtained) and use of the data is limited to the purpose for which it was collected.

Since different datasets may be used for the training, testing or validation (or similar subsets) of the AI system, such datasets should be periodically reviewed and updated. This is a particularly important issue for an AI system, whose database must be retrained, often with new data sources that have not originally been considered by the development team, and whose data may include very large number of parameters.

As discussed above, some aspects of data management are even more critical when using AI systems. These include data privacy, security and robustness. Actuaries are encouraged to consider whether privacy and security risks are addressed in alignment with enterprise policies and applicable regulations, with particular attention paid to actuarial datasets (e.g., claims histories, sensitive demographic data) where breaches could have a disproportionate impact on individuals.

#### *Data privacy*

This refers to the practice of ensuring that personal information is collected, processed, stored and shared in a way that respects an individual's rights and expectations of confidentiality. In a data-rich environment it should be noted that alternative data, alone or in combination, could identify a protected data feature whose identification may need extra effort and care to prevent. Data privacy is a key component of data protection and cybersecurity, aimed at preventing unauthorized access, misuse or exposure of sensitive information.

Key principles of data privacy include:

- **Transparency:** Clearly inform users how their data is collected, used and shared;
- **Consent:** Obtain consent to use the data;
- **Data Minimization:** Collect and retain only what is needed for the specific purpose;
- **Security:** Put in place adequate measures to protect data against breaches and unauthorized access; and
- **Accountability:** Demonstrate compliance with privacy laws and regulations, for example.

#### *Data security*

This refers to the measures and practices implemented to protect digital information from unauthorized access, corruption, theft or destruction throughout its lifecycle. Effective data security not only protects sensitive information but also:

- Ensures compliance with legal and regulatory requirements;
- Preserves organizational reputation and trust; and
- Mitigates financial losses from data breaches or cyber attacks;

Key aspects of data security are:

- **Confidentiality:** Ensuring that sensitive information is accessible only to authorized individuals or systems. This may include the need to anonymize some of the Personally Identifiable Information (PII);
- **Integrity:** Protecting the data from being altered or tampered with, ensuring its accuracy and trustworthiness; and
- **Availability:** Ensuring that the data is accessible to authorized users when needed.

#### *Data robustness*

This refers to the quality and stability of data used to train, validate and test a model. It ensures that the model performs consistently across various conditions, even when exposed to noisy or incomplete data. Features of robust data include:

- **Accuracy:** It represents real-world scenarios as precisely as possible;
- **Diversity:** Diverse scenarios, covering edge cases and anomalies;
- **Consistency:** It is uniform in structure, format and definition;
- **Completeness:** Missing data or incomplete records should be addressed appropriately;
- **Resilience to Noise:** The model should perform well despite the presence of noise in the input data; and
- **Timeliness:** It is associated with the appropriate time periods.

Data robustness can be tested and enhanced through:

- **Data Augmentation:** Introducing variations (e.g., noise, transformations) to simulate diverse conditions;
- **Outlier Detection and Handling:** Identifying and managing anomalies to prevent skewed results;
- **Use of Synthetic Data:** Generating data to fill gaps or simulate rare cases; and
- **Adversarial Testing:** For example, intentionally exposing the model to incomplete data to identify vulnerabilities.

The complexity and software-like nature of some AI systems pose a particular challenge in managing data privacy, security and robustness. As such, actuaries are encouraged to consider involving data and cybersecurity professionals at this stage, where proportionate to the risks involved.

### **3.3.2 Training and Evaluating the AI Model**

The step of training and evaluating (testing and validating) the AI model relies upon training and testing data that has been gathered and prepared as discussed in 3.3.1, using an iterative process until the model performs as intended. For example, should the AI model perform poorly on the training data, the actuary will have to improve the model, which can be achieved by selecting a better algorithm, increasing the quality of data or feeding more data to the model. Should the model not perform well on testing data, then it might fail to generalize well to unseen data.

Evaluation criteria such as performance, stability and reproducibility are consistent across most models, but validation of AI may require an actuary to adapt validation methods to the specific type of AI. This adaptation may be required for several reasons:

- AI may involve the transformation of unstructured data;
- Previously manual steps such as variable selection and dimensionality reduction may be included in an automated fashion and should be evaluated;
- In addition to out-of-sample or out-of-time validation based on test statistics, explainability may or may not be essential to the actuarial or business purpose validation of a model. This explainability may be required by internal or external parties;
- It can be difficult to evenly compare models at different levels of transparency. An actuary may need to weigh predictive power vs. explainability in the validation process; and
- The model is trained, validated, and tuned using the training and the validation data sets respectively. The main objective of this step is to ensure that the model generalizes well to unseen data upon its deployment in the real world. For this purpose, multiple experiments should be carried out on the model using all three datasets – training, validation and testing – to bring out its best abilities and minimize the changes it undergoes post-deployment.

Testing AI and ML models in actuarial work is essential due to the unique complexities and risks associated with these technologies. Unlike traditional actuarial models, which often rely on established statistical methods, AI and ML models can exhibit unpredictable behaviours and biases that may not be immediately apparent. Key elements to consider during testing include accuracy, fairness, robustness and explainability. Actuaries are encouraged to test that models not only perform well on historical data but also generalize effectively to new, unseen data. This requires a comprehensive approach to testing that includes defining clear objectives, creating diverse test cases, and ensuring high-quality data preparation to minimize bias and enhance reliability.

The role of the actuary in testing AI and ML models is multifaceted. Actuaries may be responsible for validating model performance, assessing the ethical implications of model outputs and ensuring compliance with regulatory standards while leveraging their expertise in risk management to identify potential pitfalls in model design and implementation. This involves not only evaluating the technical aspects of the models but also understanding the broader implications of their use in decision-making processes. Actuaries may also participate in the development of testing protocols and contribute to the establishment of governance frameworks that promote transparency and accountability in AI applications.

To effectively test AI and ML models, actuaries can employ a variety of techniques and metrics. Common approaches include functional testing, bias and fairness testing, and robustness testing against adversarial inputs. Metrics such as precision, recall and F1 score can be used to measure accuracy, while statistical tests can help evaluate fairness across different demographic groups. Explainability techniques, such as SHAP values, can provide insights into model decision-making processes. Continuous testing and monitoring are also crucial, as they allow actuaries to track model performance over time, identify emerging biases, and implement necessary adjustments through feedback loops and retraining protocols.

The testing of AI and ML models in actuarial work requires a rigorous and structured approach that goes beyond traditional methodologies, which is why actuaries are encouraged to continue learning about managing the risks of these models. Actuaries play a critical role in ensuring that models are reliable, fair and ethically sound. By employing a range of testing techniques and metrics, they can navigate the complexities of AI and ML, ultimately contributing to more informed and responsible decision-making in the actuarial field. For more details on testing AI and ML models, refer to the IAA paper Testing of Artificial Intelligence Models or Systems.

### 3.3.3 Documenting the AI System

Given the inherent complexity and lack of transparency of AI systems – often referred to as the “black box” effect – it is essential to document the key attributes of the model through the model’s lifecycle. Documenting the AI system is crucial whether the model is developed internally or outsourced to third parties. Consideration might be given to including a summary of key risks identified, the controls implemented to address them and how these controls will be monitored over time in the documentation.

While documenting the system, it is important to adhere to the principle of proportionality (i.e., documentation depth should scale with significance, risk and complexity), to ensure that the level of detail in the documentation is proportionate to the potential risk and implications of the AI system. The table below sets out the key attributes related to the documentation of the various elements of the AI model. For more details, refer to the IAA paper Documentation of Artificial Intelligence Models or Systems.

Elements	Attributes
Data	<ul style="list-style-type: none"> <li>• Data flow and data inventory;</li> <li>• Data quality and data governance;</li> <li>• Data usage;</li> <li>• Data limitations and weaknesses; and</li> <li>• Data privacy and security.</li> </ul>
Model development	<ul style="list-style-type: none"> <li>• Model overview;</li> <li>• Model selection, model training, and validation process;</li> <li>• Model performance, evaluation, and interpretation;</li> <li>• Model assumptions, risk, and limitations; and</li> <li>• Version control and management.</li> </ul>
Model deployment and maintenance	<ul style="list-style-type: none"> <li>• Model deployment process;</li> <li>• Monitoring and feedback loop; and</li> <li>• Model maintenance.</li> </ul>

Elements	Attributes
Other disclosures	<ul style="list-style-type: none"> <li>• Ethical considerations and legal considerations;</li> <li>• Roles and responsibility;</li> <li>• Risk assessments; and</li> <li>• Reliance on other data or models.</li> </ul>

Consideration could also be given to providing a user guide on how to use and interact with the model.

### 3.4 Approving the AI System

The formal approval of an AI system prior to its implementation and deployment is an essential part of a strong model governance framework. The type of approval required may vary based on the nature and intended use of the AI system as captured by its model risk rating. Higher risk AI systems may be subject to more rigorous review and approval processes. High-risk AI systems may be subject to independent validation (see Section 2.8), with final approval given by the appropriate governance committee once any exceptions produced as the result of the independent model validation are resolved.

Actuaries may be involved with this process as members of the development team, the independent validation team or the oversight committee responsible for the AI system’s final approval.

### 3.5 Implementing the AI System

When implementing AI systems, accountability is key to their success. Accountability can refer to different areas, such as:

- **Implementation:** The mechanism put in place to ensure who is responsible and accountable for implementing the system;
- **Operational effectiveness:** Ensuring the system is transparent and adheres to ethical standards;
- **Auditability:** Ensuring the design, data, processes and algorithms are compliant, fair and accurate; and
- **Redress:** A very important aspect when implementing AI systems, referring to the human intervention, which at times should be able to address and correct any harm and errors caused by AI systems.

Trustworthiness, compliance and accountability are essential for implementing AI systems and require adherence to best practices. DevOps methodologies were mentioned in Section 3.3 Developing the AI System, and following practice-based frameworks such as, but not limited to, Mops is important. Mops frameworks cover the entire lifecycle management of a system, including development, deployment, monitoring and retraining. By integrating DevOps practices, data engineering and machine learning algorithms, an Mops framework enhances the synergy between these domains. The CI/CD (continuous integration and continuous deployment) practices within this framework automate workflows across different

organizational segments, particularly among various stakeholders. This automation ensures that roles and responsibilities are maintained, even when using concurrent models or scaling models to embrace technological innovations.

The benefits of employing a framework such as the MLOps framework when implementing AI systems are manifold. It fosters efficiency through streamlined collaboration among diverse disciplines and people such as actuarial science, data science, IT professionals and lawyers. Automation of various stages – ranging from data preparation and feature engineering to training, testing, deployment and monitoring – minimizes manual errors. Furthermore, it enables standardization of processes across different lifecycle stages of a model, ensuring consistency. Scalability is another advantage as it allows AI systems to be dynamically adjusted based on business needs for enhanced performance. Lastly, compliance is ensured, as the AI systems implemented adhere to the legal and regulatory requirements specific to their application region.

In addition to establishing a system management framework, it is essential to integrate the system into the organization's culture. Furthermore, as outlined in Section 2.1 Roles and Responsibilities, it is crucial to clearly define and thoroughly document the involvement and motivation of various stakeholders during the implementation phase. Moreover, the primary obstacle to digital transformation is often the lack of know-how. Actuaries are encouraged to facilitate the interconnection of various disciplines and stakeholders, particularly when certain parties are less accountable for the system itself. This can be achieved by organizing workshops, creating straightforward documentation and providing cross-functional training. Cross-functional training is an effective method for improving a team's skills, collaboration and performance. It allows team members to experience different roles, functions and perspectives within the organization, and to learn from one another.

Given AI's scalability and adaptive nature, it is important to implement standards, best practices and quality tools from the start. Actuaries are encouraged to take timely action to ensure AI systems comply with applicable professional standards and legal and regulatory requirements, to mitigate industry-specific regulatory risks and liabilities towards model users in case of defects.

### **3.6 Ongoing Monitoring of the AI System**

Continuous model monitoring and retraining are essential practices to maintain the performance, reliability and relevance of AI models, especially as data distributions shift or new information becomes available. AI systems inherit biases from the environment they operate in, therefore not directly accounting for potential harm. Thresholds should be set for model performance metrics and drifts, and alerts should be triggered.

Such thresholds need to account for well-defined Key Performance Indicators (KPIs), business objectives, and a set of parameters that can be easily adapted to address improvements on the data and model quality. Once a model is being deployed into a well-established framework, such as the MLOps mentioned in Section 3.5, monitoring the performance and accuracy of the model and its alignment with the business goals becomes necessary. Due to the potentially higher risk of model drift in AI models – caused by data drift, for example – it is important to consider the following points:

- Continuously monitor model performance for fitness for purpose, including monitoring user inputs and outputs to identify potential unexpected results that compromise the fitness for purpose;
- Define parameters and thresholds (e.g., grid search, optimization, fairness criteria) that adapt and help in measuring model performance and accuracy at any future iteration point in time;
- Create an environment to monitor training and test datasets, changing data environments and changing business goals to be able to address model drifts, deficiencies or potential harm; and
- Use the environment and monitoring as a function to redress the model, should the model be negatively changed by its environment.

In addition, one needs to be continuously concerned about the changes in model performance and data robustness (see Section 3.3.1 for more details). If variability or unforeseen changes occur, those aspects can be monitored. Model performance measures how effectively an AI system achieves its objectives based on specific metrics. It reflects the model's ability to generalize from training data to unseen scenarios. Some key metrics and monitoring elements to use when evaluating model performance are:

- Accuracy: The proportion of correctly predicted instances;
- Data and model drifts, which relate to the data and model alteration during time, such as changing the distribution of features and new patterns, on which models become unable to perform appropriately;
- Precision and Recall: Relevant for imbalanced datasets where false positives or false negatives matter. False positives in data science are the results that indicate that a condition exists, while it actually does not exist. False negatives are the opposite: a test result indicating a condition does not exist, when it exists;
- Quantitative goodness-of-fit validation on an independent dataset for AI-driven regression models;
- Automation and proactive notifications when dealing with performance issues; and
- Monitoring of memory consumption, latency and GPU/CPU usage.

Actuaries are encouraged to test that all these aspects are considered and fed into audit trails and testing, and are implemented in a stable technical environment. Moreover, once the processes of an AI system can be monitored, an automated governance framework should be established to ensure that regulatory requirements and conformity assessments are completed in a timely manner before a model is deployed.

Having a strong foundation for AI system monitoring also enables actuaries and organizations to decommission models more effectively. When retiring an AI system or model, the metrics and overviews gathered provide valuable insights. Clear performance assessments, safety reviews, fairness tests and financial viability tests help communicate findings to stakeholders. They also facilitate post-mortem analysis, offering lessons for future processes, promoting knowledge transfer among teams and mitigating any residual risks arising from retired systems.

## 4. Additional Considerations

### 4.1 Training and Education

To build an organization-wide culture of AI governance, training at different organizational levels is essential. Bridging the gap between disciplines and functions is important, as strengths and weaknesses are unavoidable. Teams responsible for the use of AI should have received appropriate training to reduce the likelihood of deployment errors.

Executive Leadership (C-Suite, board):

- Goal: Ensure strategic alignment, risk management and oversight.
- Focus: Ethical AI, risk and compliance, strategic goals and accountability.
- Format: Executive briefings and workshops with case studies.

Senior Management (department heads, AI leaders):

- Goal: Embed governance practices within departments.
- Focus: Operationalizing AI governance, bias mitigation, resource needs and risk management.
- Format: In-depth workshops and scenario simulations.

AI Practitioners and Data Scientists:

- Goal: Equip with skills to design, test and monitor governed models.
- Focus: Bias detection, data privacy, explainability, robustness testing and documentation.
- Format: Live practice sessions, coding workshops and toolkits.

IT and Operations Teams:

- Goal: Consistent governance in deployment and monitoring.
- Focus: Model lifecycle, data quality, monitoring and security.
- Format: Training in MLOps, data governance and technical workshops.

Legal, Compliance and Risk Management:

- Goal: Oversee compliance and risk.
- Focus: Regulatory knowledge, risk assessment, data privacy and audits.
- Format: Legal briefings and scenario-based compliance training.

All Employees (awareness training):

- Goal: Foster awareness of AI ethics and governance.
- Focus: AI basics, ethical implications, data security and reporting mechanisms.
- Format: E-learning modules, webinars and ethics workshops.

Ongoing Education:

- Offer certification programs, refresher courses, and industry seminars to keep skills and knowledge up to date.

This structured training ensures AI governance is practised effectively and aligns with organizational goals across all levels.

## **5. Conclusion**

The presented AI governance framework breaks governance into a few relevant parts – data, training, validation, implementation and human involvement throughout. Each step requires interpretation to be applied to various forms of AI, depending on its complexity and application. Because of this interpretation, AI training, as mentioned in Section 4, should not be undervalued or overlooked. Knowledge of AI will develop a culture of appropriate skepticism – and enthusiasm – that is needed to implement a governance framework around these revolutionary technologies.

AI is constantly evolving, and the governance frameworks for AI should continue to evolve with it. We expect this paper to be updated as new research becomes available, notably with respect to interpretability and bias-related research.

Actuaries are well suited for governing AI, because the actuarial skillset has traditionally focused on both rigorous statistical models and the application of actuarial judgment to account for how models may misbehave when applied directly in practice. The introduction of AI into actuarial practice may make this a more complicated endeavour, but not an unfamiliar one. We hope that this paper helps actuaries translate their existing skillsets into the new world of AI governance.

## Appendix: Definitions – Bias, Fairness and Discrimination

### 1. Introduction

The increasing use of advanced analytic techniques in actuarial sectors (whether this be predictive analytics, machine learning, data science, AI in primarily numerical contexts or Generative AI in typically non-numerical contexts) has led many parties in these sectors (actuaries, actuarial bodies, regulators, risk and/or compliance officers, policyholder and/or consumer champions, politicians, trade commentators, etc.) to express concerns about *bias*, *fairness* and *discrimination*.

However, these terms are generally used very loosely, to the extent that artificial barriers to AI adoption are created through misunderstanding.

It is important that both practitioners and stakeholders are clear on, and do not confuse, the meaning of these terms. Most of the actuarial services, in particular insurance and pensions, are risk-differentiation activities, and this activity, and the associated use of AI to assist in this, should not be presumed to be discriminatory in its prejudicial sense.

The terms “bias”, “fairness” and “discrimination” should not be used interchangeably.

It is important to emphasize that the descriptions and definitions given below are, to some extent, designed to capture the meaning of *bias*, *fairness* and *discrimination* when it comes to actuarial services in AI systems. Although some parts of the descriptions and definitions may remain valid outside this area, considerations should be given to their validity in other areas.

### 2. Summary and Definitions

The definitions of *bias*, *fairness* and *discrimination* below are deliberately short and simple; the next section provides more depth, and these high-level definitions should not be used without regard to the context and market. These definitions have been established for the purpose of this paper and are not formal IAA definitions. They have been arrived at from discussions among members of the IAA AITF rather than being taken verbatim from specific reference sources, although many sources have been considered and utilized for the special purpose as described above.

The extent of variation across different contexts and markets, as well as the extent of misunderstanding about the meaning of these words, is such that *we suggest actuaries define what they mean by these terms when they use them*.

- *Bias*: Bias means some input, output or other element of the AI system is not representative of what is regarded as reality.
- *Fairness*: Fairness relates to an individual or group being treated reasonably given their reasonable expectations.
- *Discrimination*: Discrimination is the unfair treatment of an individual or group relative to others.

### 3. More detailed considerations

This section expands on the deliberately brief definitions provided above.

#### 3.1 Bias

Modern-day applications of big data and AI have generated more scrutiny of the meaning of the term “bias” than at any time since its first usage in the 16th century. Adopted from the French word “biais”, the term has evolved from meaning “slant”, “oblique”, “diagonal” or “against the grain” to a legal meaning of “undue propensity or prejudice”, a statistical meaning of systematic error, and a meaning adopted by the social sciences as a preference for one subpopulation over another. While bias may not result in an adverse impact, it can, and because of a hyperfocus on the adverse impacts of bias the term has been imbued with negative connotations.

Many standard-setting organizations have developed definitions of the term in response to the algorithmic outcomes that are perceived to unfairly disadvantage some and advantage others. For example, the OECD contends that algorithms that advantage some populations compared to others in regard to characteristics such as gender, race and ethnicity, migration status and so on can be labeled as labelled bias. Further, the OECD contends that such bias occurs as a result of societal biases that are discriminatory towards certain groups and become encoded in algorithmic outcomes.<sup>12</sup> The OECD acknowledges other sources of bias, such as statistical and human, but did not explicitly include them in its definition of bias.

The National Institute of Standards and Technology (NIST), rather than devising a definition of bias, recognized the definition devised by the International Organization for Standardization (ISO), which states that bias is the degree by which a reference value deviates from the truth.<sup>13</sup> This definition of bias may be deemed problematic because it requires an assertion of truth against which a deviation can be measured. It may be problematic, as such an assumption may itself reflect bias.

In statistics, there is the notion of an unbiased estimator that measures the deviation between the expected value of a sample statistic of a parameter and its true population.<sup>14</sup> If the deviation is zero, then the bias is zero.

A sample statistic may be distorted by, for example, sampling and measurement biases that would result in its deviation from its theoretical true value. Other types of statistical biases can result in distortion in statistical metrics as well, including selection bias, temporal bias and outlier bias. Statistical biases lend themselves to quantification and correction. For example, a statistically unbalanced, non-representative sample can be discarded for a more

---

<sup>12</sup> OECD (2023), *OECD Digital Education Outlook 2023*, [https://www.oecd.org/en/publications/oecd-digital-education-outlook-2023\\_c74f03de-en/full-report/algorithmic-bias-the-state-of-the-situation-and-policy-recommendations\\_a0b7cec1.html](https://www.oecd.org/en/publications/oecd-digital-education-outlook-2023_c74f03de-en/full-report/algorithmic-bias-the-state-of-the-situation-and-policy-recommendations_a0b7cec1.html)

<sup>13</sup> ISO (2006), *ISO 3534-1:2006 Statistics – Vocabulary and Symbols, Part 1: General Statistical Terms and Terms Used in Probability*, <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/04/01/40145.html>

<sup>14</sup> This relationship is formulated as

$$\text{Bias} = E[\hat{\theta}] - \theta,$$

where the expected value of a sample statistic of a parameter ( $\hat{\theta}$ ) and its true population of ( $\theta$ )

representative one by resampling. However, not all biases can be captured by formulaic representation. The approach taken by NIST addresses this shortcoming.

NIST developed a taxonomy of bias that reflects three mutually distinct categories that cover most if not all the types of biases that can impact AI models: statistical, human or cognitive, and systemic. The statistical category is defined as discussed above and is purely a quantitative and not qualitative approach to assessing bias. NIST recognizes that detecting and addressing only statistical biases is not sufficient to eliminate harm in AI resulting from cognitive and social systemic sources of bias. Detecting bias from these sources requires socio-technical approaches where societal policies and practices are considered and examined, as well as human behaviours that may influence the generation of modelling data.

For example, beliefs regarding the existence of climate change and extreme weather events could result in the collection of data that only supports that belief. The data collected would be factually correct, and a purely statistical approach would not detect the human and cognitive behaviours that influenced how the modelling data was generated. Similarly, reflecting only historical data and weather patterns in catastrophe models may result in the same myopic results and lead to incorrect predictions from AI models, with consequences for decision-making.

Further examples of data bias can be seen where readily available datasets are used that are not properly representative of the subject matter (such as insurance policyholder or pensioner subsets). Even cleaning and adjusting data in a way that is intended to improve the dataset can introduce slight biases (e.g., eliminating data where some fields are missing or thought likely to be incorrect). The decision as to what to do with data collected during the coronavirus pandemic is a good example of how bias may be introduced by *not* making substantial adjustments to the data (e.g., in a mortality dataset, not excluding the high mortality periods of 2020/21 or making some other adjustment in respect of COVID-19 deaths).

As NIST would contend, these examples demonstrate that models may be statistically correct but still produce biased results, and these biased results may create harm. A full discussion of cognitive and societal biases is beyond the scope of this paper, but these two categories have implications for the responsibility of the actuary in interrogating modelling data and AI model outcomes for these two categories of bias.

As the preceding discussion suggests, it is not enough to rely on statistical approaches alone to detect, mitigate or eliminate bias in AI models. Understanding human or cognitive biases, which number into the hundreds, is a necessary aptitude to detect when biases such as confirmation bias, groupthink bias and anchoring bias can result in non-representative outcomes.

Systemic biases are harder to detect and arrest than the other two because they require knowledge of societal policies and practices, and the groups that are adversely impacted by them. From just looking at a dataset, the many decisions that resulted in shortcomings in the modelling data or the design of the AI model may not be obvious. For example, a health algorithm resulted in people of colour being referred to specialized care less often than white patients. The target variable in the model was health care spending, a variable that many believed to be innocuous, but the model resulted in adverse health consequences for people of colour. It was not immediately obvious to the modellers who built the model that health care spending is less in communities of colour because of a lack of access to healthcare, distrust of the medical establishment, and unequal treatment rooted in medical racism or the

different societal embeddedness of colour due to different socializations that come with birth and go back to generations inheriting the trans-generation patterns that maintain cultural attitudes. The mathematics underlying the algorithm were sound, but the algorithm still produced harmfully biased results.

The definitions and examples discussed illustrate that an overarching definition of bias should be broad enough to be true for every type of bias that may influence AI outcomes. The term “algorithmic bias” is problematic because it is an umbrella term that masks the many ways in which human biases can influence the design, evaluation and implementation of algorithms. For instance, using a Generalized Linear Model to analyze risk without allowing for material interactions between factors is likely to mis-quantify the risk attaching to individuals who exhibit two or more relatively extreme data characteristics (e.g., a very high or low age and a very high or low benefit amount). Algorithms can amplify biases already present in the data.

Harmful biases are the main concern of regulators and may require socio-technical skills to detect. The social science definition of bias generalizes well to cover statistical, human or cognitive, and systemic biases. However, it is unlikely that only one type of bias will pervade a given AI result. Context is important. Rather than rely on an overarching definition of bias, actuaries are encouraged to identify, analyze and appropriately treat the specific biases in the specific context that result in harmful bias in the outcome of AI models and AI systems. Actuaries may need training in socio-technical approaches to adequately detect, mitigate or eliminate harmful biases in AI models used in their work.

Note that in certain circumstances bias may not be harmful or inappropriate. For example, an implicit safety margin on an estimate of an expected value of a random variable would be a biased estimate of the true expected value but its use is intentional and not harmful/inappropriate.

### **3.2 Fairness**

While fairness appears to be an intuitively simple concept, relating to the reasonable treatment of people, there is ample scope for misunderstandings to arise in typical actuarial contexts.

The potential for misunderstanding becomes clear if we start with what we might call “actuarial fairness in insurance pricing”: that an individual’s premium (or, more generally, the relationship between benefits received and premiums paid by them or on their behalf) reflects their level of risk.

There are many concrete definitions of (actuarial) fairness; all of them have some advantages and disadvantages compared to the others. No single definition has emerged as the best one in literature.

While insurance essentially offers protection against volatility risks (an individual usually pays a fixed or increasing premium pattern related to the expected cost of a claim, rather than “self-insuring” and retaining exposure to a sudden large claim), many also view it as risk-sharing between heterogeneous groups.

The minimal level of risk-sharing can be seen in medically underwritten policies, where an individual’s premium-to-benefit relationship is set on the basis of age, sex (European Union [EU] and ex-EU countries excepted), and various medical data points such as body mass index, medical history and so on. Typically, genetic information is excluded from the underwriting process and thus the only risk-sharing (ignoring volatility risk) relates to protected characteristics (e.g., race in many countries) and any hereditary disorders.

At the other extreme, in some markets it is normal for premiums in a life insurance context to be set with regard only to age (and perhaps any major medical history), leading to sharing across genders and the smoker/non-smoker divide.

Many might regard the above sharing as “fair”, while from an actuarial perspective in a life insurance context women would be subsidizing men and non-smokers subsidizing smokers.

The sharing can also be conceptually extended from the notion of groups of heterogeneous individuals to the idea of smoothing risks over time. For instance, with-profits business has in many markets smoothed investment return fluctuations via bonus mechanisms; some policyholders will have gained from this, while others will have lost.

So, what is regarded as fair depends on:

- a. Legal norms; and
- b. General societal norms.

An insurer or pension provider will decide where it wants to fit within those. Some insurers will offer more granular pricing, some will not.

Therefore, any definition of fairness necessarily involves what people can reasonably expect, based on the country/market they are in, and what their insurer or pension provider has communicated to them. There is no objective “fair” or “unfair”.

The fact that risk-sharing can be imposed “at the stroke of a pen” also illuminates the arbitrary nature of the concept and the lack of any universal, objective fairness – for example, the EU introducing gender-neutral insurance pricing in the early 2010s.

Finally, note that fairness may, and often does, in common usage encompass notions of legal equality, respect for individuals, consistency and transparency. The last two aspects are particularly important to bear in mind in contexts where it is likely that individuals may think they have been unfairly treated; these are also aspects that are more of a challenge in an AI context than in traditional actuarial work.

### **3.3 Discrimination**

The meaning of the word “discrimination” is widely addressed in various sources. As a result, there are slightly different definitions:

- The unjust or prejudicial treatment of different categories of people, especially on the grounds of ethnicity, age, sex, or disability (Oxford);
- Treating a person or particular group of people differently, especially in a worse way than the way in which you treat other people, because of their race, gender, sexuality, etc. (Cambridge);
- Prejudiced or prejudicial outlook, action, or treatment (Merriam-Webster); and
- Discrimination is the process of making unfair or prejudicial distinctions between people based on the groups, classes, or other categories to which they belong or are perceived to belong, such as race, gender, age, class, religion, or sexual orientation (Wikipedia).

The EU adopted several non-discrimination directives that apply the following types of definitions:

- Direct discrimination: Where one person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of a protected characteristic; and
- Indirect discrimination: Where an apparently neutral provision, criterion or practice would put persons of a protected characteristic at a particular disadvantage compared with other persons unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary.

Usually, different types of discrimination are distinguished:

- *Age discrimination* (ageism) is a type of discrimination based on one's age.
- *Gender discrimination* refers to the unequal or disadvantageous treatment of an individual based on their gender.
- *Disability discrimination* (ableism) occurs when a person is treated unfairly or less favourably due to their physical or mental disability.
- *Genetic discrimination* refers to the unequal treatment of individuals based on an aspect of their genetic code or genome, such as the risk for genetic disorder. Genetic discrimination can involve such genomic information being used against individuals in a variety of circumstances, such as employment, health or disability, insurance status, education or health care.
- *Racial discrimination* occurs when an individual is subjected to unequal treatment or harassment because of their colour, or their actual or perceived race. Policies or practices that have a disproportionately negative impact on members of a particular race may be considered discriminatory, even if they appear neutral.

Other commonly recognized forms of discrimination relate to:

- Military status;
- National origin;
- Religion or belief;
- Being pregnant or on maternity leave; and
- Sexual orientation.

Finally, there may be instances of "de facto" discrimination which do not relate to any of these well-recognized categories. For instance, in some markets attention is now being given to "vulnerable customers" to ensure that insurers (in particular) do not in effect discriminate against the less wealthy and/or less educated members of the public by virtue of costly products, complex policy wording and so on.

To be contrasted with *differentiation*, where individuals are assessed according to existing laws/regulations, *discrimination* relates to some feature that is prohibited under the relevant laws/regulations or wider societal/market practice.

The definitions and examples discussed illustrate that an overarching definition of discrimination should be broad enough to be true for every type of discrimination that may be caused or amplified by the application of an AI system within the jurisdiction concerned.

Actuaries are usually familiar with the concept of differentiation; for example, in the case where policyholders are assessed according to their riskiness, which leads to sufficient individual premiums during the pricing process. In contrast, the different types of possible discrimination that could be caused by the application of an AI system depend heavily on the input data and the algorithm used. A particular challenge is indirect discrimination relating to the use of a proxy (e.g., gender might be rated indirectly via occupational categories – nursing being a female-dominated occupation compared with construction work being a male-dominated occupation).

The European Commission issued an instructive Communication<sup>15</sup> that sheds light on the thin line between allowed differentiation and not-allowed indirect discrimination when using certain characteristics as rating factors that may exhibit correlation with gender but are valid rating factors in their own right (see Points 16 and 17 and Footnote 3 of that document in particular).

Measurement of and managing the related risks are one of the future demanding challenges for actuaries deploying or using AI systems.

---

<sup>15</sup> European Commission (2012), Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats), [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012XC0113\(01\)](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012XC0113(01))

## References

European Commission (2012), Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats),

[https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012XC0113\(01\)](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012XC0113(01))

European Commission (2019), Ethics guidelines for trustworthy AI,

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

European Union Agency for Fundamental Rights (2022), Bias in Algorithms: Artificial Intelligence and Discrimination,

[https://fra.europa.eu/sites/default/files/fra\\_uploads/fra-2022-bias-in-algorithms\\_en.pdf](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf)

International Actuarial Association (2018), International Standard of Actuarial Practice 1 General Actuarial Practice,

[https://actuaries.org/actuarial\\_practices/isap-1-general-actuarial-practice/](https://actuaries.org/actuarial_practices/isap-1-general-actuarial-practice/)

ISO (2006), ISO 3534-1:2006 Statistics – Vocabulary and Symbols, Part 1: General Statistical Terms and Terms Used in Probability,

<https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/04/01/40145.html>

Monitoring Authority of Singapore (2018), Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector,

<https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/feat>

Organisation for Economic Co-operation and Development (2023), OECD Digital Education Outlook 2023,

[https://www.oecd.org/en/publications/oecd-digital-education-outlook-2023\\_c74f03de-en/full-report/algorithmic-bias-the-state-of-the-situation-and-policy-recommendations\\_a0b7cec1.html](https://www.oecd.org/en/publications/oecd-digital-education-outlook-2023_c74f03de-en/full-report/algorithmic-bias-the-state-of-the-situation-and-policy-recommendations_a0b7cec1.html)

Organisation for Economic Co-operation and Development (2024), Explanatory Memorandum on the Updated OECD Definition of an AI System,

[https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system\\_3c815e51/623da898-en.pdf](https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf)

Organisation for Economic Co-operation and Development (2024), Recommendation of the Council on OECD Legal Instruments Artificial Intelligence,

<https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>