



**International Actuarial Association
Association Actuarielle Internationale**

Testing of Artificial Intelligence Models or Systems

**AI Task Force
November 2025**

IAA Paper

Testing of Artificial Intelligence Models or Systems

This paper was prepared by the Artificial Intelligence Task Force (AI Task Force) 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
605 - 75 Albert St, Ottawa ON K1P 5E7 Canada

www.actuaries.org

© International Actuarial Association/
Association Actuarielle Internationale

Table of Contents

1. Introduction	1
2. Setting the foundations for Testing	2
2.1 Identify Objectives and Requirements	2
2.2 Identify Test Cases and Scenarios	3
2.3 Establishing a Testing Environment	3
3. Preparing Data for Testing	3
3.1 Data Collection and Preparation for Testing	4
3.2 Data Augmentation Techniques	5
3.3 Data Labelling and Annotation	5
3.4 Data Privacy and Security Considerations	6
4. Testing the Key Requirements of Trustworthy AI Models	6
4.1 Testing Metrics	7
4.2 Accuracy	7
4.3 Fairness	8
4.4 Explainability	9
4.5 Robustness	10
5. Fairness and Ethical Considerations	10
5.1 Data Bias and Fairness	10
5.2 Responsible Disclosure of Vulnerabilities	11
5.3 Governance and Monitoring	12
6. Types of Testing	12
6.1 Functional Testing	13
6.2 Integration Testing	13
6.3 Output Appropriateness Testing	13
6.4 Security Testing	14
6.5 Bias and Fairness Testing	14
6.6 Explainability and Interpretability Testing	14
7. Continuous Testing and Monitoring	15
7.1 Continuous Testing	15
7.2 Continuous Monitoring	15
7.3 Feedback Loops and Model Re-Training	16
7.4 Model Versioning and Post-Deployment	17
8. Conclusion	17

1. Introduction

In recent years, the integration of artificial intelligence (AI) and machine learning (ML) into actuarial practice has been reshaping the landscape of risk assessment and decision-making. As these technologies become increasingly prevalent, it is important to understand the principles and methodologies involved in creating, testing and using AI models to ensure their reliability, accuracy and fairness. Actuaries play a critical role in this process, leveraging their expertise in statistical analysis, risk assessment and regulatory compliance. Through comprehensive testing of AI models, actuaries help mitigate model risk and enhance the overall integrity of the decision-making process. This paper aims to provide educational material regarding good practices for testing AI models within the context of actuarial work, emphasizing a principle-based approach that allows for flexibility in application. By presenting suggested approaches and tools for effective model testing, this paper aims to foster confidence in the use of AI technologies while promoting ethical and responsible practices.

The paper is structured to offer an overview of the key considerations and methodologies involved in testing AI models. The structure of the paper is set out below. Some points may appear in more than one section as they are relevant in the different contexts. Additionally, the paper addresses the importance of model governance and risk management, highlighting the actuary's role in the testing and validation process.

- Section 2 sets out the foundation of testing which includes defining objectives and requirements around core principles (fairness, explainability, accountability, transparency, robustness), identifying test cases and scenarios and establishing a testing environment such as infrastructure and tools
- Section 3 introduces practical approaches to assemble high-quality, representative, and well-documented datasets—typically larger and more diverse than for traditional actuarial tests—to improve reliability, fairness, and regulatory alignment
- Section 4 explains approaches to testing the core principles of ethical AI models – accuracy, fairness, explainability, robustness—and emphasizes standardized metrics to evidence performance and support governance and continuous improvement
- Section 5 considers wider fairness and ethical measures such as detecting and mitigating data bias, disclosing vulnerabilities and implementing governance to oversee ethical deployment and continuous monitoring
- Section 6 summarizes the various methods of testing, each of which focuses on a particular aspect of the model

Section 7 describes a lifecycle approach including continuous testing and monitoring, feedback loops and model re-training after testing, model versioning post deployment. This paper is intended to be read alongside the Artificial Intelligence Governance Framework paper.¹ It provides supplementary details specifically focused on the testing of AI models within the broader AI governance context. The intended readers of this paper are actuaries and professionals involved in actuarial work who are seeking to enhance their understanding of AI model testing. In this document the terms “AI system” and “AI model” align with those presented in the Governance Framework paper. In summary, an AI system is an overall machine-based system that infers from inputs how to generate outputs, and an AI model is a core component within an AI system used to make such inferences.

¹ IAA Paper: Artificial Intelligence Governance Framework

https://actuaries.org/app/uploads/2025/11/AITF_Governance_Framework_Paper_Final_Approved.pdf

2. Setting the foundations for Testing

2.1 Identify Objectives and Requirements

The first step in preparing for testing AI systems is to clearly define the objectives and requirements of the testing process. This involves understanding the specific goals that the AI system aims to achieve and the criteria for success.

The testing objectives and requirements depend on a variety of factors, including (but not limited to) the:

- Purpose of the system
- Operator of the system
- End user of the results
- Quality of the required data
- Complexity of the model or system
- Risk level of the model or system
- Impact to customers
- Knowledge of AI models
- Company's IT environment

Additionally, the testing objectives and requirements should satisfy the following principles:

- Fairness
- Explainability
- Accountability
- Transparency
- Functionality (performance, security, usability, compatibility)
- Robustness
- Trust

For actuarial work, AI systems may be adopted for a wide variety of purposes, ranging from regulatory reporting to pricing and underwriting, and so on. Notwithstanding the variety of factors outlined above, it is recommended that actuaries verify that the corresponding principles are applied consistently, taking into account applicable local regulations and guidelines. Where AI systems are adopted to produce a customer outcome, customers should be treated fairly.

It is important to document the testing objectives, environment and requirements comprehensively, as they will guide the entire testing process and ensure that all stakeholders have a clear understanding of what the testing aims to accomplish (see the Documentation of Artificial Intelligence Models or Systems paper² for good practices on the documentation of AI systems).

² IAA Paper: Documentation of Artificial Intelligence Models or Systems

https://actuaries.org/app/uploads/2025/11/AITF_Documentation_AI_Models_Paper_Final_Approved.pdf

2.2 Identify Test Cases and Scenarios

Following the definition of the testing objectives and requirements, the next step is to identify relevant test cases and pass criteria. This involves creating a diverse set of scenarios that the AI system is likely to encounter in production environment applications.

It is important that the test cases cover a wide range of possibilities to thoroughly evaluate the AI system's performance. Testing aims to strengthen the AI system so that it produces consistent and reliable results across the intended scope of population. For example, test cases could assess whether there could be persons or groups disproportionately affected by negative performance of the AI system.

Considerations should be given to testing "Boundary" or "tail-end" scenarios, as AI algorithms may not produce sensible results, possibly due to a lower volume of training and/or validation data.

Testing standards issued by local regulators or actuarial professional bodies on testing actuarial systems should be adhered to, where appropriate.

2.3 Establishing a Testing Environment

Establishing a controlled and consistent testing environment is important for evaluating the AI system's performance. This involves setting up the necessary infrastructure, tools and frameworks to conduct the tests. Many organizations maintain internal procedures, datasets and testing frameworks that are used in system development, updates and testing and should be considered.

It is also important that the environment is secure and that data privacy is maintained throughout the testing process.

Testing of AI systems may require a large amount of computing power and time (in addition to a large amount of data). For example, a Large Language Model (LLM) may require a high amount of computing power and number of iterations to process the vast amount of testing data required. In such cases, additional computing capacity is necessary to meet the testing requirements. This is to ensure that the AI system is thoroughly evaluated and ready for deployment in the production environment.

Stakeholders should consider incorporating AI ethics into the AI assessment implemented in their organization, which can be integrated into the testing environment.

3. Preparing Data for Testing

Well-prepared datasets help to increase the reliability and robustness of AI system testing. From data selection to data labelling to addressing data biases, each aspect plays a key role in the quality and suitability of testing datasets for evaluating the performance of an AI system. Having good-quality datasets for testing also helps to address potential biases and enhance model integrity, which are important elements of ethical AI.

Additionally, maintaining data quality, integrity and comprehensive documentation is a practice that supports robust model testing and compliance with regulatory standards, including actuarial standards on data.

This section introduces practical approaches and actuarial considerations regarding producing high-quality testing datasets.

3.1 Data Collection and Preparation for Testing

AI system testing typically requires more data than traditional software or actuarial system testing due to the need for extensive training, testing and, in many cases, validation datasets. Large and diverse datasets are very useful for supporting the model's robustness and generalizability.

Data collection and preparation are critical components of the testing process. The quality and relevance of the data used for testing will significantly impact the accuracy and reliability of the testing outcome. Testing datasets need to be representative of the population or scenarios for which the AI system will be deployed. This involves considering the characteristics of the data (demographic, geographic, temporal, etc.). Incorporating diverse data sources is important to capture a wide range of scenarios. Apart from internal data (e.g. a policy admin system, claims records, finance data), external data could be used, such as publicly available data or industry-specific data sources, to supplement the test data.

To achieve accurate representation, stratified sampling techniques could be used. They involve dividing the dataset into subgroups based on relevant characteristics and then sampling from each subgroup proportionally. This helps prevent underrepresentation or bias in the testing data. The corresponding features within each subgroup that are critical for the model's performance should be identified to ensure they are well represented.

The data needs to be representative of the population and scenarios that the AI system could encounter. The data should be gathered from reliable sources, and it needs to be clean, accurate and free from biases (for more details, see Section 5, Fairness and Ethical Considerations). In some actuarial use cases, certain proxy data fields indirectly reflect characteristics associated with certain groups of individuals. Examples include postcode, occupation, income level, education level, marital status and religious beliefs. It is important to assess whether the data introduces unintended bias or discrimination against any groups of individuals.

Common data preparation issues and solutions are applicable in the context of AI system testing. For example, data preparation may include normalizing data, handling missing values and formatting data correctly for the AI system. To the extent applicable, data standards should be applied on data used for testing AI systems.

It is recommended that the quality of the testing datasets is assessed; for example, by checking for missing values, outliers, extreme values or any other data issues that may affect the model's performance. Strong controls – facilitated by automation and data analytics – would help to improve data quality.

Data preparation for AI systems may involve specific procedures not normally required for traditional actuarial systems; for example:

- Applying In-Sample/Out-of-Sample/Out-of-Time schema to partition data into training, testing and validation datasets
- Labelling data in different groups for a classification algorithm
- Balancing the data to manage data imbalance issues (e.g. through under- or over-sampling, applying class weights)
- Data encoding to transform data from one form to another (e.g. categorical values to numerical values)
- Feature selection for a machine learning model

In such cases, actuaries could seek advice from experts in the relevant fields and adopt the appropriate data preparation techniques.

Furthermore, testing of AI systems may require a large amount of data. For example, a large language model may require a vast amount of testing data to ensure that sufficient test cases are covered. This amount of data may be challenging to generate out of conventional actuarial processes. In such cases, the actuary could seek advice or additional datasets to support the testing requirements.

3.2 Data Augmentation Techniques

Data augmentation techniques can enhance the variety and size of the testing datasets. Common techniques include synthetic data generation, over-sampling, under-sampling and noise injection.

- *Synthetic data generation* – This involves creating artificial samples that closely resemble real data using techniques such as Generative Adversarial Networks (GANs). This can be useful when real data is limited or difficult to obtain.
- *Over-sampling* – This involves increasing the representation of minority classes in the test data to address class imbalance issues.
- *Under-sampling* – This involves reducing the representation of majority classes in the test data to address class imbalance, particularly when the majority class overwhelms the model's ability to learn from the minority class.
- *Noise injection* – This involves adding random variations to the test data to introduce robustness and generalize the model's performance.

It is important to ensure that the augmented data maintains the integrity and relevance of the original data and does not introduce unintended biases or distortions. Consideration should be given to using statistical methods where possible to substantiate the augmented data in terms of relevance, representation and fairness.

3.3 Data Labelling and Annotation

Accurate and reliable labels are valuable for training and testing AI systems. Clear guidelines and processes for data labelling and annotation are valuable to provide consistency and minimize labelling biases.

Guidelines provide detailed instructions on how to label different data types and ensure that annotators have a clear understanding of the desired outcomes. Quality control measures could be implemented, such as inter-rater reliability checks, to verify the correctness and reliability of labels. Multiple annotators can independently label the same data to assess consistency and resolve discrepancies.

Having a "human-in-the-loop" to provide oversight helps to increase confidence in data labelling. If necessary, expert knowledge in or an expert review of the labelling of certain types of data, such as medical or legal data, could be considered. Involving domain experts in the annotation process can enhance the accuracy and relevance of the labels.

Regular reviews of the labelling process should be conducted to identify and address any issues that may arise, such as evolving labelling guidelines or changes in annotator performance. The review process should be part of the model governance framework.

3.4 Data Privacy and Security Considerations

Data privacy and security considerations are important when preparing, creating, updating, and testing systems and data. This applies to the handling of both the original data and data generated and stored within the AI system itself. Actuaries are encouraged to adhere to data minimization principles, collecting and using only the data necessary for model functionality. In the case of health data, for example, it is advisable to avoid using personally identifiable information (PII) unless absolutely necessary. Instead, actuaries can implement unique identifiers to track participants without exposing their identities.

To enhance data privacy, techniques such as anonymization and pseudonymization can be considered for this purpose.. This can be achieved through methods like k -anonymity or differential privacy, which mask individual data points. Additionally, implementing strict access controls ensures that only authorized personnel can access sensitive information.

In addition to protecting individual data points, care should be taken to avoid privacy leakage through grouped or aggregated outputs. Even when individual identifiers are removed, small or unique subgroups can be unintentionally exposed through statistics or visualizations. This risk should be assessed when designing test datasets and outputs, and monitoring dashboards.

It is important to maintain clear documentation of data collection processes, consent forms, and privacy considerations surrounding data usage. Legal advice may be required when considering data privacy. Conducting periodic privacy impact assessments can further enhance data protection measures and instil public confidence in the handling of sensitive information. This involves mapping data flows and identifying where personal data is used in testing. As part of model risk management, privacy impact assessments could be performed to identify and mitigate potential privacy risks associated with the use of testing data. This assessment considers factors such as data storage, transfer and sharing processes.

Other practical steps to protect data include involving data protection officers from the outset to ensure that all testing activities comply with legal requirements and regular training for staff on data protection regulations to maintain compliance and uphold ethical standards. Compliance on data privacy regulations (e.g. the General Data Protection Regulation, or GDPR, in Europe) must be followed when handling testing data, particularly when personal data is involved. This means ensuring that PII is properly anonymized or de-identified to protect individual privacy for usage.

Encryption and access controls should be implemented to safeguard sensitive information and prevent unauthorized access. The data preparation processes should adhere to industry-standard encryption protocols and establish strict access protocols for testing datasets.

Informed consent should be obtained from individuals whose data is being used for testing purposes. The necessary consent and opt-out opportunities should have been obtained and communicated to the individuals whose data is intended to be used for testing purposes.

4. Testing the Key Requirements of Trustworthy AI Models

Testing plays a crucial role in supporting the requirements of reliability, accuracy, fairness and performance of AI models. The outcome of robust testing is that these requirements are sufficiently tested in terms of both coverage and rigor, including the use of relevant metrics to substantiate the testing outcomes.

When AI models are adopted for actuarial work, actuaries need to be aware of the testing requirements necessary for these complex models, which may require a different approach

to that for testing conventional actuarial models (be it deterministic or stochastic). It is important to demonstrate that, as part of model governance, the AI models are sufficiently tested.

The importance of rigorous testing increases when the model outcome has an impact on decisions that concern customers directly (especially, if there are no humans in the loop). Examples include (but are not limited to):

- Underwriting and premium setting
- Surplus distribution and bonus setting on participating contracts
- Claims processing and payment
- Assumptions setting

The extent to which testing is performed may vary based on the model's intended use, taking proportionality into consideration (i.e. the testing depth should reflect risk, complexity and materiality). For example, when a model is used to support underwriting decisions, particular attention may be given to evaluating fairness. The testing standard for a model used to inform an underwriter will likely be different from a model used to independently make underwriting decisions.

4.1 Testing Metrics

Testing metrics are standardized and measurable benchmarks for evaluating the key aspects of the model such as fairness, performance and robustness. They play an important role in providing objective measures and insights into the AI model's functionality, enabling model developers to compare and assess different models or iterations, and making it easier to identify areas for improvement, address biases or vulnerabilities and make evidence-based decisions.

From a governance perspective, the use of testing metrics provides key stakeholders and regulators with tangible and verifiable evidence of the model's capabilities. Additionally, the use of metrics allows for continuous monitoring and improvement, helping the model to remain robust, reliable and fit for purpose over time.

The rest of this section sets out the key requirements of the AI models that should be tested with relevance for actuarial work, some possible metrics to substantiate testing outcomes and key actuarial considerations.

4.2 Accuracy

Overview: To assess accuracy, the model's predictions could be compared with known ground truth labels, results from previous systems and observed experience. This can be achieved by dividing the dataset into training and testing sets, where the testing set contains labelled data that the model has not seen before to test whether the model generalizes well to unseen data.

Testing Approach: Cross-validation techniques such as k -fold cross-validation could be used. This process involves splitting the dataset into k subsets, training the model on $k-1$ subsets, and validating it on the remaining subset. This is repeated k times, with each subset used once as the validation data. Additionally, comparing model predictions against a known ground truth helps in assessing the model's correctness. Independent peer review can further strengthen confidence in the model's results.

Testing Metrics: The following metrics could be used to measure model accuracy in the context of supervised learning:

- *Precision and Recall* – These measure the model’s ability to correctly identify positive instances and its sensitivity to false negatives, respectively.
- *F1-Score* – This is a harmonic mean of precision and recall that provides a balance between the two.
- *Mean Absolute Error (MAE) and Mean Squared Error (MSE)* – These quantify the average prediction error.

For probabilistic models, the actuary is encouraged to consider scoring rules such as the Continuous Ranked Probability Score (CRPS) and the Energy Score, which assess the full predictive distribution rather than point predictions.

Key actuarial considerations are:

- *Relevance* – The data used for training and testing should provide useful historical information, and also be relevant for future scenarios in which the model is deployed. Sources of data may be internal (e.g. policy data, claims data) or external (e.g. national statistics, financial market data). The ideal data sources vary depending on the purpose of the model.
- *Data partitioning* – Having a sufficiently large dataset is important for model testing (and training, for that matter), as it can be partitioned into training and testing data subsets. Approaches including taking an “out-of-time” hold-out sample (the most recent period of the data set that has been kept independent of the model training process in order to validate performance), a random sample or a more complex selection of a data subset. Depending on the use case, having a large enough dataset may be a challenge; for example, while building an underwriting model for a relatively new life insurance product, when available and relevant data is scarce. The risk of introducing bias into testing data should also be managed.

4.3 Fairness

Overview: Fairness testing involves evaluating model outputs across different demographic or socio-economic groups to identify potential biases. Techniques such as re-weighting or re-sampling can be used to mitigate bias.

Testing Approach: The fairness of an AI model can be evaluated by measuring and comparing the model’s predictions and their outcomes for different demographic or socio-economic subgroups. Statistical methods, such as calculating false positive rates and false negative rates, and examining group disparities in predictions, can uncover potential biases.

Testing Metrics: The following fairness or statistical metrics could be used to test fairness:

- *Disparate Impact Ratio* – This measures the ratio of favorable outcomes between different subgroups.
- *Equal Opportunity Difference* – This assesses the difference in true positive rates across subgroups.
- *Equalized Odds* – This ensures that the probability of true positive and false negative is the same across subgroups.

Key actuarial considerations are:

- *Fairness definition* – There are many definitions of fairness, depending on the industry and model use case; which one to use is often a business decision, reflecting the goals and outcomes of the specific business area. Businesses may find it challenging to

discuss fair outcomes, and may not have a clear philosophy on fair outcomes compared to fair process. This challenge is best addressed at a higher level as part of the risk management framework or in corporate policies regarding the business concerned; for example, principles and practices of managing a with-profits fund.

- *Sources of bias* – Biases may be introduced in various ways, such as through the data used to train AI models, algorithms written in the model, inputs from third-party models, interpretation of the model results, human biases and feedback bias flowing back to the model as inputs. Differentiation can be intended (e.g. higher insurance premiums for higher-risk occupations) and may be justified, and can also be unintended, which should be avoided. The testing framework should distinguish between “good” and “bad” biases. In the past, actuaries may have placed less attention on testing fairness for traditional actuarial models, and so model testing guidelines may need to be enhanced to provide guidance on testing fairness, especially if the model outcome has a direct and financial impact on policyholders.

See Section 5, Fairness and Ethical Considerations, for further details. A discussion on bias, fairness and discrimination is included in the Appendix to the Artificial Intelligence Governance Framework paper.

4.4 Explainability

Explainability will differ significantly from one type of AI model to another. Some models will have parameters that are easily represented using existing techniques. These 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. However, this approach may not be applicable to other AI models. Truly black-box models will require the interpretation of abstract scores and indexes, and it is recommended that the evaluation of explainability of these models is commensurate with their usage. Where a high level of explainability is needed, some AI models may be inappropriate for use without a clear understanding of the risks associated.

Overview: Testing the explainability of a model involves assessing its interpretability and the extent to which it can provide explanations for its decisions or predictions. This can be done by analyzing the model’s internal mechanisms, such as feature importance, to understand how it arrived at specific outputs.

Testing Approach: Various approaches to assess explainability include (depending on the type of AI model):

- Feature importance analysis to rank the key drivers for model prediction
- SHAP values to understand the marginal effect of each feature on the target variable
- Area Under the Receiver Operating Characteristic curve (AUC-ROC curve) to graphically understand the performance of a binary classification model

Testing Metrics: The following metrics could be used to test explainability:

- *Feature Importance Rank* – This identifies which features most influence predictions.
- *Shapley Value* – This assesses the impact of a feature on model output.
- *AUC Score* – This evaluates the discriminative power of a binary classification model.

Key actuarial considerations are:

- *Focus on the customers* – Approaches to test and manage explainability issues should not only focus on model features and parameters (treating this purely as a modelling exercise), but also on customer outcome and communications. This also means that “explanation” requires context. For example, the requirement for a good external explanation for a customer or regulator is different from that for internal explanations for modellers.
- *Consult AI/ML experts* – Explainability and the associated testing metrics may be new to actuaries. In such a case, consider seeking advice from the relevant experts during the testing process to properly design the guidelines and test cases. Remediation approaches should be considered.

4.5 Robustness

Overview: Robustness testing involves introducing noise or adversarial examples to evaluate the model’s resilience to various types of perturbations, adversarial attacks or out-of-distribution inputs.

Testing Approach: This may include injecting noise into the input data, manipulating key features or introducing adversarial examples. Stress testing under different scenarios, such as varying data distributions, helps assess performance consistency.

Testing Metrics: The following metrics could be used to test model robustness:

- *Robustness Score* – This quantifies the model’s resilience small, often deliberate, changes to a model's inputs or parameters.
- *Sensitivity Analysis* – This examines how changes in input affect outputs.
- *Performance Degradation Rate* – This measures the decline in performance under stress conditions.

Key actuarial considerations involve:

- *Existing guidance and practices* – Testing how models behave under stress conditions or react to outlier data is not new to actuaries, because this is an established practice for testing actuarial models. Hence, existing guidance and practices are relevant and should be adopted to the extent appropriate. However, it should be noted that AI models may not be suitable for “edge cases” by nature, and so appropriate testing should indicate where the “boundaries” (and hence limitations) are.

5. Fairness and Ethical Considerations

As actuaries increasingly incorporate AI and ML into their work, it becomes important to address the fairness and ethical considerations when testing the models. This section outlines key fairness and ethical considerations, along with practical examples and good practices, during the testing process.

5.1 Data Bias and Fairness

One of the foremost concerns in the testing of AI systems is the need to support fairness and mitigate bias. The testing process itself should be designed to be free from bias, so that evaluations do not favour any particular group or outcome.

Actuaries are encouraged to be vigilant about how test data selection might introduce bias. For instance, if a model predicting insurance claim approvals reveals that older adults consistently receive lower approval rates relative to younger individuals, this may indicate an age bias that requires rectification.

Testing datasets should be analysed for potential biases by assessing the representation of different demographic or socio-economic groups or examining the correlation between input features and protected attributes.

Diverse datasets could be used to reflect various demographic groups and scenarios. Datasets could also include real-time data to sufficiently test for bias and fairness in real-time production environments.

Actuarial work may sometimes involve a wide variety of personal data. Actuaries need to be mindful that proxy bias could arise where training data includes variables that act as proxies for protected attributes. Even though these protected attributes may not be explicitly included in the data, the AI system may indirectly infer them from proxy variables, leading to biased outcomes. Test cases should cover testing for proxy biases.

Practical approaches to mitigate biases include conducting fairness audits to assess the impact of testing data biases on model outputs. This involves comparing the performance of the model across different demographic or socio-economic sub-groups to identify and rectify any disparities.

Rigorous audits of the datasets used in model testing involve employing statistical methodologies and bias detection measures, such as analysing the distribution of outputs across different demographic groups. Data audits can be facilitated using external or internal tools to quantify fairness gaps. If biases are detected, data augmentation techniques to balance the representation of underrepresented groups could be considered.

If biased features are identified in the test data, the implications and potential consequences of using such data should be considered, especially against regulatory requirements on meeting policyholders' reasonable expectations and treating customers fairly.

5.2 Responsible Disclosure of Vulnerabilities

Responsible disclosure of vulnerabilities is another important aspect of ethical considerations in model testing. Regularly conducting assessments to identify potential vulnerabilities in the decision-making processes of the underlying models should be considered. For instance, discovering that a model is overly sensitive to specific inputs could lead to unintended consequences, such as an excessive number of claims being denied due to anomalous feature values.

Establishing a clear protocol for reporting vulnerabilities is valuable, encompassing steps for internal assessments and external notifications when necessary. This could involve setting up a process for reporting and addressing vulnerabilities; for example, something similar to a bug bounty program.

Collaboration with cybersecurity teams is crucial for ensuring that system vulnerabilities are promptly identified and mitigated. Whenever feasible, publishing findings related to model limitations promotes transparency and fosters trust with the public.

Fostering a culture of openness and continuous improvement encourages team members to report potential issues without fear of retribution. Collaborating with external experts for review and validation, by engaging third-party auditors, can provide independent verification of results.

5.3 Governance and Monitoring

The establishment of accountability and governance frameworks supports ethical AI system testing. It is important to define clear roles and responsibilities so that specific individuals or teams are accountable for ethical decision-making in testing. This could involve the formation of an ethical review board or ethics committee to oversee AI deployments and maintain ethical standards. These committees can provide guidance on ethical dilemmas and support compliance with industry standards. Alternatively, the scope of the Internal Audit Committee could be extended to cover AI principles.

Establishing clear, unbiased criteria for evaluating model performance is also crucial. This might involve setting thresholds for acceptable error rates across different groups.

Engaging stakeholders, including community representatives, is valuable for understanding external perceptions of fairness and validating the model's impacts. Involving diverse teams in the testing process can provide multiple perspectives, further reducing the risk of bias and enhancing the overall ethical standing of the model.

Continuous monitoring and improvement of AI systems are important to support their continued ethical application. Establishing systems for ongoing performance tracking of models post-deployment allows actuaries to identify shifts in performance or fairness. For instance, automated monitoring tools can alert actuaries if performance metrics deviate significantly, permitting timely interventions.

6. Types of Testing

It is recommended that testing of the AI system is conducted as early as possible to support reliability, fairness and security throughout the development and validation process, as well as after development.

It is recommended that all components of an AI system are included in testing:

- *Data* – Ensuring data quality, consistency and relevance is crucial, as AI systems depend heavily on the data they are trained on.
- *Models* – Testing models for accuracy, robustness and generalizability ensures they perform well on unseen data and under various conditions.
- *Environments* – The deployment environment must be tested to ensure compatibility and performance under real-world conditions.
- *The system as a whole* – End-to-end testing ensures that all components, including infrastructure (e.g. GPUs (graphics processing units)), work together seamlessly, providing the desired outcomes.

For the purposes of this paper, the focus is on the testing of the models.

Testing methods applied on conventional software systems or actuarial models may be insufficient for testing AI systems due to the complex, non-deterministic and context-specific nature of AI algorithms:

- *Complexity* – AI systems often involve complex algorithms and large datasets, making traditional testing methods inadequate for capturing all potential issues.
- *Dynamic behaviour* – AI models can change behaviour based on new data (e.g. machine learning models), requiring continuous testing and validation.

- *Non-deterministic outputs* – Unlike traditional software, AI systems may produce different outputs for the same input, necessitating specialized testing approaches to ensure consistency and reliability.

The key testing paradigms performed on AI systems are outlined below.

6.1 Functional Testing

Functional testing checks whether the AI system performs its intended functions correctly. This involves verifying the accuracy of predictions, classifications or recommendations made by the AI system. Test cases should cover a wide range of inputs, including edge cases and scenarios that the model might encounter in real-world applications.

AI systems often involve probabilistic/stochastic models that may produce different outputs for the same input due to inherent randomness or learning dynamics. Probabilistic/stochastic models are not new to actuarial work; hence, actuarial standards for testing stochastic models could be applied.

While the principles of functional testing are similar to the testing applied to non-AI actuarial models, the complexity of AI systems means that actuaries need to understand the underlying AI algorithms to create effective test cases. Also, the dynamic behaviour of AI systems means that continuous monitoring and testing may be required as the model evolves with new data.

6.2 Integration Testing

Integration testing focuses on the interactions between different components of the AI system. This includes verifying that data flows correctly between data sources, preprocessing modules, models and output interfaces. Integration testing ensures that the system works cohesively and that changes in one component do not adversely affect others.

While actuarial systems testing would test that different actuarial models and data sources integrate correctly, AI systems involve more complex data pipelines, including data preprocessing, feature engineering and model serving. Components such as APIs (application programming interface) should be tested for stability as part of integration testing.

Furthermore, dynamic datasets that reflect real-time data flows and interactions between components are required. This may not be standard practice for testing actuarial systems; hence, building a robust data platform is important for not only successful AI system testing but also deployment.

6.3 Output Appropriateness Testing

Output from the model, including the user interface, should be checked for appropriateness to ensure that the results/visualizations are clear and presented in an easy-to-interpret manner, and can be actioned upon correctly.

Dynamic visualizations involving the data and results are important for AI systems. This may not be the case for traditional actuarial systems where the priorities are normally accuracy, reliability and auditability, and so on. Testing on how complex AI outputs are presented in an intuitive and actionable manner for the end user would be a valuable part of testing.

Feedback from end users, such as actuaries and analysts, is crucial for refining the system's output in terms of content and format. Good practice is to involve the end users in the design and execution of output appropriateness testing.

6.4 Security Testing

Security testing identifies vulnerabilities and tests that the AI system is protected against threats such as data breaches, unauthorized access and adversarial attacks. This includes testing for secure data handling, encryption and access controls.

Given the potentially sensitive nature of actuarial data (e.g. policyholder data), robust security measures are important and should be thoroughly tested to ensure that sensitive actuarial data is protected.

Testing should also include malicious inputs to the AI system to assess how it reacts and whether the security measures (e.g. data encryption, access rights, threats scanning) are effective. Since actuaries are often not experts in security testing, it is recommended that relevant experts are consulted in performing such testing.

6.5 Bias and Fairness Testing

Bias and fairness testing assesses the AI system for potential biases in data and model predictions (A discussion on bias, fairness and discrimination is included in the Appendix to the Artificial Intelligence Governance Framework paper³). This involves analysing the system's outputs across different demographic groups to ensure that it does not unfairly disadvantage any group. Techniques such as fairness metrics, bias detection algorithms and diverse training datasets are employed to mitigate bias.

For fairness testing, actuaries could consider defining and applying appropriate fairness metrics as testing criteria. This may come from existing actuarial literature,⁴ such as fairness to participating policyholders.

6.6 Explainability and Interpretability Testing

Explainability and interpretability testing assesses whether the AI system's decisions and predictions can be understood and trusted by users. This is particularly important in actuarial work, where transparency and accountability are key considerations.

Such testing is only applicable to models that are fully transparent (such as GLM (generalized linear model) and GAM (generalized additive model) or explainable models (such as GBM (gradient boosting machines) and clustering), where changes to characteristics can be explained in terms of general behaviour and/or effect of the change. In these cases, techniques, such as model-agnostic explanations and feature importance analysis, and visualization tools help make the AI system's behaviour more interpretable. Clear documentation of test cases and their outcomes contribute to overall transparency and may align with existing actuarial guidance on model testing and documentation. For models that are completely black box (such as ChatGPT), explainability remains limited given the current maturity of AI technology, though efforts to improve this are being made. In contexts where the rationale behind model outputs cannot be meaningfully understood or validated, the use of such models may not be appropriate within actuarial work.

³ https://actuaries.org/app/uploads/2025/11/AITF_Governance_Framework_Paper_Final_Approved.pdf

⁴ <https://www.casact.org/publications-research/research/research-paper-series-race-and-insurance-pricing>

7. Continuous Testing and Monitoring

Continuous testing and monitoring of AI systems plays an important role in maintaining performance, reliability and compliance with regulatory requirements. This section sets out key considerations for continuous testing and monitoring of AI systems, specifically from an actuarial perspective.

7.1 Continuous Testing

Continuous testing involves setting up a systematic and ongoing process to assess the accuracy, robustness, fairness and compliance of AI models. Actuaries are encouraged to consider the following:

- *Test Coverage* – A comprehensive set of test cases is required to cover different scenarios, input data distributions and edge cases. These tests should evaluate the model's performance across various aspects, such as accuracy, reliability and risk assessment. For example, in developing a climate risk model, test cases could simulate extreme weather events to test that the model can produce sensible results to these challenging conditions.
- *Test Data* – The test data used needs to represent real-world scenarios, including historical data and simulated data for future scenarios. The data should be diverse, inclusive and free from bias to avoid perpetuating unfair outcomes. For example, a pricing model could consider market shifts to evaluate how the model responds to changing economic conditions.
- *Performance Metrics* – Appropriate performance metrics should be established to measure the model's accuracy, reliability and fairness. These metrics should align with the objectives of the specific actuarial task and be regularly assessed to identify any deviations or performance degradation; for example, AUC (Area under the Curve) for imbalanced datasets.
- *Test Automation* – Automated testing procedures enable efficient and scalable testing. The use of testing frameworks, tools and libraries specifically designed for AI model evaluation should be considered. Examples include automated testing frameworks like TensorFlow Model Analysis (TFMA), or writing test cases within the code as part of an error-handling mechanism.
- *Documentation* – It is important to maintain comprehensive documentation of the testing process, including test cases, test data, performance metrics, and any observed issues or anomalies. This documentation should be easily accessible to facilitate scrutiny during audits and compliance reviews. (see the Documentation of Artificial Intelligence Models or Systems paper⁵ for good practices on the documentation of AI systems))

7.2 Continuous Monitoring

Continuous monitoring is necessary to ensure models continue to perform as intended, remain compliant with regulations, and detect potential biases or issues. Actuaries are encouraged to consider the following:

⁵ IAA Paper: Documentation of Artificial Intelligence Models or Systems
https://actuaries.org/app/uploads/2025/11/AITF_Documentation_AI_Models_Paper_Final_Approved.pdf

- *Real-time Data Monitoring* – Real-time dashboarding tools such as Grafana or Kibana help to visualize incoming data and outputs for evaluation purposes. This includes monitoring data quality, identifying data drift (e.g. changes in the statistical distribution of inputs, such as a sudden increase in a particular age group) and concept drift (e.g. changes in the relationship between inputs and outputs, such as revised underwriting rules), and tracking potential biases introduced during model operation. For instance, in an underwriting model, monitoring tools can track the distribution of inputs to detect data drift before it affects the model's accuracy. In addition, real-time monitoring should include alert mechanisms, audit logging and integration with MLOps (machine learning operations) pipelines to ensure prompt investigation and remediation when anomalies or security incidents occur.
- *Model Performance Tracking* – Metrics and mechanisms to monitor the model's performance post-deployment need to be established. This includes monitoring accuracy, reliability and fairness over time, and comparing the model's outputs with the expected outcomes.
- *Model Explainability* – AI model explainability allows users to interpret and understand the factors influencing the model's predictions. It enables them to detect potential biases or errors and take appropriate corrective actions. For instance, in an income protection underwriting model, understanding how different factors (e.g. occupation, gender, age, postcode) contribute to decision-making helps in identifying and mitigating any potential biases.
- *Feedback and Issue Reporting* – Feedback loops and reporting mechanisms are required to capture feedback from users, stakeholders and other impacted parties. This feedback helps in identifying potential issues, biases or unintended consequences, enabling timely model adjustments and improvements.

7.3 Feedback Loops and Model Re-Training

Feedback loops play an important role in continuously improving AI models. Actuaries are encouraged to consider the following:

- *Establishing Feedback Channels* – Develop channels for stakeholders to provide feedback on the model's performance and predictions, or any observed biases. This may include customer feedback forms, user surveys or regular interactions with impacted parties.
- *Data Augmentation and Model Updates* – Use feedback, including identified biases or errors, to update the training dataset by augmenting it with new data or modifying existing data. This process helps improve the model's accuracy, fairness and relevancy. For example, if patterns of discrimination are identified against a subset of applicants, actuaries can adjust the dataset to include more representative samples, ensuring a fairer model output.
- *Iterative Model Retraining* – Actuaries could define protocols for periodic model retraining based on feedback and data updates. Iterative retraining ensures that the model remains up to date, adaptive to changing scenarios, and aligned with evolving regulations and business requirements. Such protocols could be established, as part of model risk management, annually or less frequently, depending on the riskiness of the model, the frequency of new data and regulatory changes.

7.4 Model Versioning and Post-Deployment

Model versioning is valuable for tracking model changes, maintaining transparency and ensuring accountability. Actuaries are encouraged to consider the following:

- *Version Control* – Implement version control mechanisms to track changes made to AI models, including modifications to algorithms, training data or hyperparameters. This facilitates traceability, reproducing previous results, and identifying the source of any issues or biases. Tools such as Git can be employed to track changes in model algorithms and datasets. Each model version can have a clear change log that outlines what modifications were made, and why.
- *Documentation and Reporting* – Maintain detailed documentation for each model version, including the rationale for changes, testing results, changes in training data and performance metrics. Such documentation enhances transparency, compliance and audits. Existing actuarial standards on documentation and reporting serve as a useful starting point, offering principles that should be followed and further tailored to the specific AI modelling methodology.
- *Retiring Outdated Models* – Establish processes to retire outdated or deprecated models. This involves assessing the impact of transitioning from one model version to another, mitigating risks, and ensuring a smooth transition for stakeholders and where necessary communication to affected users.

8. Conclusion

Actuaries play a critical role in the testing of AI systems and models during their whole lifecycle, leveraging their expertise in statistical analysis, risk assessment and regulatory compliance. Through comprehensive testing of AI models, actuaries help mitigate model risk and enhance the overall integrity of the decision-making process. The content of this paper provides educational material for good practices for testing AI systems and models by considering fundamental principles of model testing, and specific challenges and opportunities presented by AI/ML technologies.