Artificial Intelligence: Testing AI Systems

This document aims to help developers and deployers of artificial intelligence (AI) systems, including deployers of AI systems procured from a third-party, to test and validate that the AI system continues to work as intended over the course of the AI system's life cycle.

Note: Deployers of third-party AI systems are still accountable for the AI systems they deploy. The testing criteria and evidence available in this document can be used to test third-party AI systems deployed by your organization and should be carried out to the extent that you have the capability to do so.

Table of Contents

TrustArc

Why Test AI Systems

AI systems must undergo rigorous testing and validation to ensure they produce reliable and accurate results. AI testing helps organizations identify errors, biases, and other issues in the AI system that may impact its reliability and efficacy and improves the overall transparency and interpretability of the AI system.

Challenges

The following highlight the most common challenges to AI testing:[1]

Non-deterministic
AI systems can show different behaviors for the same input. Example: since the AI system is constantly learning, it may provide a different response even when the same information or input is used.

Lack of Testing Data
A sufficiently large amount of data is needed to accurately represent real-world scenarios that the AI system will encounter. It is not always possible, or legal, for organizations to retain large quantities of data for training or testing purposes.

Bias
Testing for bias requires a thorough understanding of the training data and the potential sources of bias. A diverse group of human reviewers is required to recognize that outcomes may be biased towards certain individuals or groups of individuals.

Interpretability
It can be extremely difficult to extract specific attributes (e.g., if the AI is used to categorize images, it may not be possible to determine what caused an AI system to recognize or categorize an image the way it did).

Sustained Testing
Traditional software testing does not need to be repeated until such time as the software is modified. However, the ever-evolving nature of AI systems (e.g., from learning and being retrained) means they need to be continuously tested.

Timing and Frequency of Testing

AI testing should be done regularly. The criteria for testing can be time-based (e.g., annually, or biannually) or event-based (before deployment, when new criteria/inputs are added), or testing can be prompted where substantial modifications to the AI system or its operating environment take place.

[1] Jeevan Bhushetty, May 25, 2023, A Complete Guide to Testing AI and ML Applications, QED42, https://www.qed42.com/insights/a-complete-guide-to-testing-ai-and-ml-applications

Types of Testing

Due to the distinct and specific features of each AI system, there is no common, unified solution to testing AI. Each AI system will require its own unique approach to testing. The best approach is to combine traditional software testing methods with specialized AI testing techniques[2], including the following:

Functional Testing

Functional testing tests the core functionality of the AI system by assessing whether its functions and features perform as intended (e.g., functional testing of a chatbot verifies that it can understand user queries and provide appropriate responses).

Usability Testing

Usability testing focuses on the user experience. It evaluates how easily users can interact with the AI system, the degree of user-friendliness, and ease and convenience during interaction. (e.g., usability testing of a conversational AI system assesses the system's understanding of human or natural language).

Performance Testing

Performance testing evaluates how well the AI system performs under different workloads. It measures factors like response time, throughput, responsiveness, scalability, and resource usage.

Security Testing

Security testing evaluates how well the AI system, and its configuration, can prevent the leakage of data it processes.

Validation Methods

Cross-Validation

Cross-validation is a statistical method used in AI performance testing. It involves dividing a dataset into smaller subsets, training the model on one of those subsets, and then testing the model against the other subsets. The process is then repeated multiple times to ensure robustness.

Comparison or A/B Testing

This method of testing is most used in AI applications to compare two versions of an AI system (e.g., to determine which version leads to a more favorable outcome).

Validation Against Ground Truth

Validation against ground truth testing involves validating the AI system against a ground truth (e.g., AI algorithms used in the medical field can be validated against diagnoses made by medical experts).

[2] Karyna Kosynova, January 16, 2024, Testing AI Applications: Best Practices and Case Study, MobiDev, https://mobidev.biz/blog/how-to-test-ai-ml-applications-chatbots

Testing AI-based Systems

AI testing should take a risk-based approach; testing should be sufficiently rigorous and commensurate with the level of risk posed by processing from the AI (i.e., whether it is used to make predictions or decisions that have a legal or similarly significant effect on individuals or provides more of an automated processing function).

Preparation

- Clearly define the objectives, scope, and success criteria of your AI testing efforts. Determine which specific aspects of the AI system will be tested (e.g., accuracy, performance, robustness, fairness)
    - Consider testing against the AI principles to which your organization commits to follow
- Create test data sets to determine the efficacy of your AI system:
    - The data sets should be logically constructed to test all possible combinations and arrangements of the data
    - Include a variety of test scenarios that represent the range of processing the AI system will face in production
- Establish baseline metrics or results to compare the AI system's performance test results against (this creates a benchmark against which the AI system can be compared)
- Where appropriate for your organization and the complexity of your AI system, consider leveraging AI-powered software testing tools

Below is an example of criteria that may be used to test against AI principles. Evidence that the AI system meets the criteria can be provided through documentation, including documentation of test results where outputs are compared, and other means that confirm that a control is in place.

Transparency

Criteria: Appropriate information is provided to individuals on the use of AI and the AI system, including at the moment an individual interacts with the AI system:

- End users/data subjects are informed about the collection/use of their personal data and risks of decisions made by the AI system
- Information provided is clear and made accessible through traditional public-facing methods (e.g., privacy notices, privacy policy), and upon request

Evidence:

- Privacy notices and privacy policies include explanations about the collection/use of personal data for AI purposes, and the risks of decisions made by AI systems
- There is a documented trail that shows privacy notices and privacy policies have been reviewed by relevant stakeholders (e.g., legal, compliance, business unit)
- Information is visible and accessible at the time an individual interacts with the AI system (e.g., through a privacy notice, pop-up message, etc.)
- There are documented processes for responding to individuals' requests for information

Explainability

Criteria: The AI system's functions and results are accurate and explainable:

- The organization can explain the factors and criteria affecting the AI system's predictions/results/decisions
- End users relying on the AI system to enable or carry out a decision understand the decision and can explain it
- Explanations are tailored to the stage of the AI lifecycle, the target audience, and the risk level of the AI system's processing
- The organization understands and can explain the provenance of the data (where it originated from, how it was collected and curated)
- The organization can trace the rules or operations that led to the AI system results

Evidence:

- There are documented explanations of the factors and criteria affecting the AI system's predictions/results/decisions which have been reviewed by relevant stakeholders (e.g., legal, compliance, business unit)
- Documented training materials and records of completion exist to show that end users relying on the system to enable or carry out a decision have been trained on the AI system and their knowledge of its outputs tested
- Different versions of explanation have been documented for different audiences based on their expected level of understanding
- Documented explanations include an explanation of the data used in the decision, where it came from and how it was collected and is used by the organization
- Documented explanations explain how the AI system uses the data or inputs to generate a decision or output (e.g., it calculates A and B to arrive at C)

Safety

Criteria: An impact assessment has been conducted

- The organization has identified and assessed risks, risk metrics and risk levels of the AI system for each specific use case, and has mitigated such risks
- The organization has identified residual risk that cannot be mitigated

Evidence:

- The organization has documented the risks, risk metrics and risk levels of the AI system for each specific use case
- Documentation includes a risk value for each identified risk (e.g., high/medium/low, probable/improbable, etc.)
- Mitigating controls are documented for each identified risk
- Risks that have no mitigating controls have been documented

Security

Criteria: Security measures are in place to protect the AI system from unauthorized access, disclosure, modification, destruction, or disruption

- Requests made to live AI systems are monitored to detect suspicious activity
- Actions (i.e., by humans and automated) involving the AI system are logged and retained
- API calls and/or input queries are monitored

- Internal limits are placed on the number of queries allowed from the same IP or with similar inputs
- Effective authentication and access controls have been implemented to mitigate inference attacks
- Actions are taken when suspicious activity or incidents occur - possible actions include to:
    - Flag the activity for review
    - Limit or block further requests
    - Suspend user accounts

Evidence:

- There are documented procedures in place to monitor live AI systems and review activities that are suspicious
- There is a record or log of actions involving AI systems
- There is a documented mechanism in place to track API calls and/or input queries and documented procedures to monitor them
- There is a documented mechanism and logic in place that limits the number of queries made to the AI system from the same source or with similar inputs
- Access controls are documented and in place to limit access to the AI system to those who have the organization's permission
- Documented procedures are in place that describe the actions to be taken when different types of suspicious activities or incidents occur

Robustness

Criteria: The AI system can continue to function even when unexpected inputs are introduced

- AI system results are repeatable; they can be consistently replicated or reproduced
- Changes in the AI system's performance is analyzed to determine how robust the AI system is

Evidence:

- Documentation exists showing that the AI system can generate the same or significantly equivalent results when new circumstances are introduced, such as changes to the operating environment or presence of other factors (human or artificial) that may interact with the system in an adversarial manner

Fairness

Criteria: AI system results do not result in unintended and inappropriate discrimination against individuals or groups

- The AI system makes the same decision even if an attribute is changed
- Fairness metrics are continuously monitored to ensure they do not violate predetermined thresholds

Evidence:

- There is a documented comparison of results when two diverse groups are tested based on the negative and positive values used (e.g., gender groups, class groups, age groups)
- There are documented metrics to judge the fairness of AI system outcomes
- There are documented procedures to continuously monitor and assess AI system outputs against documented fairness metrics

Human oversight and control

Criteria:

- Human operators are appropriately trained and equipped with the necessary tools, information, and authority to provide proper oversight, and have the technical ability to intervene in decision-making processes where necessary

Evidence:

- There are documented training procedures for human operators; completion records are documented
- Training procedures for human operators documents their authority to intervene in decision-making processes

Report Test Findings

Test results are often presented as a statistic where validation of the AI algorithm yields range-based accuracy or confidence scores rather than expected outcomes.

Resources

NIST AI Risk Management
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Jeevan Bhushetty, May 25, 2023, A Complete Guide to Testing AI and ML Applications, QED42,
https://www.qed42.com/insights/a-complete-guide-to-testing-ai-and-ml-applications

AI Verify Foundation AI Governance Testing Framework and Toolkit – Sample Report
https://aiverifyfoundation.sg/downloads/AI_Verify_Sample_Report.pdf

Karyna Kosynova, January 16, 2024, Testing AI Applications: Best Practices and Case Study, MobiDev,
https://mobidev.biz/blog/how-to-test-ai-ml-applications-chatbots