Microsoft Responsible AI Standard, v2

GENERAL REQUIREMENTS

FOR EXTERNAL RELEASE

June 2022

Microsoft

Index

About this release

When we embarked on our effort to operationalize Microsoft's six AI principles, we knew there was a policy gap. Laws and norms had not caught up with AI's unique risks or society's needs. Yet, our product development teams needed concrete and actionable guidance as to what our principles meant and how they could uphold them. We leveraged the expertise on our research, policy, and engineering teams to develop guidance on how to fill that gap.

The Responsible AI Standard is the product of a multi-year effort to define product development requirements for responsible AI. We are making available this second version of the Responsible AI Standard to share what we have learned, invite feedback from others, and contribute to the discussion about building better norms and practices around AI.

While our Standard is an important step in Microsoft's responsible AI journey, it is just one step. As we make progress with implementation, we expect to encounter challenges that require us to pause, reflect, and adjust. Our Standard will remain a living document, evolving to address new research, technologies, laws, and learnings from within and outside the company.

There is a rich and active global dialog about how to create principled and actionable norms to ensure organizations develop and deploy AI responsibly. We have benefited from this discussion and will continue to contribute to it. We believe that industry, academia, civil society, and government need to collaborate to advance the state-of-the-art and learn from one another. Together, we need to answer open research questions, close measurement gaps, and design new practices, patterns, resources, and tools.

As we continue our journey, we welcome feedback on our approach and insights on other ways forward: https://aka.ms/ResponsibleAIQuestions

Accountability Goals

Goal A1: Impact assessment

Microsoft AI systems are assessed using Impact Assessments.

Applies to: All AI systems.

Requirements

A1.1 Assess the impact of the system on people, organizations, and society by completing an Impact Assessment early in the system's development, typically when defining the product vision and requirements. Document the effort using the Impact Assessment template provided by the Office of Responsible AI.
Tags: Impact Assessment.

A1.2 Review the completed Impact Assessment with the reviewers identified according to your organization's compliance process before development starts. Secure all required approvals from those reviewers.
Tags: Impact Assessment.

A1.3 Update and review the Impact Assessment at least annually, when new intended uses are added, and before advancing to a new release stage.
Tags: Impact Assessment.

Goal A2: Oversight of significant adverse impacts

Microsoft AI systems are reviewed to identify systems that may have a significant adverse impact on people, organizations, and society, and additional oversight and requirements are applied to those systems.

Applies to: All AI systems.

Requirements

A2.1 Review defined Restricted Uses to determine whether the system meets the definition of any Restricted Use. If it does, document this in the Impact Assessment, and follow the requirements for the Restricted Use.
Tags: Impact Assessment.

A2.2 Answer prompts in the Impact Assessment template to determine whether the system meets the definition of a Sensitive Use. If it does, report it to the Office of Responsible AI, and follow any additional requirements resulting from a Sensitive Uses review.
Tags: Impact Assessment.

A2.3 Review your systems at least annually against the definitions for Sensitive Uses and Restricted Uses. If there are systems that meet the criteria for Sensitive Uses, report them to the Office of Responsible AI. If there are systems that meet the criteria for Restricted Uses, notify the Office of Responsible AI.

Goal A3: Fit for purpose

Microsoft AI systems are fit for purpose in the sense that they provide valid solutions for the problems they are designed to solve.

Applies to: All AI systems.

Requirements

A3.1 Document in the Impact Assessment how the system's use will solve the problem posed by each intended use, recognizing that there may be multiple valid ways in which to solve the problem.
Tags: Impact Assessment.

A3.2 Define and document for each model in the AI system:
1) the model's *proposed inputs* and how well they represent the concepts they are intended to represent; include analysis of the limitations of this representation,
2) the model's *proposed output* and how well it represents the concept it is intended to represent; include analysis of the limitations of this representation, and
3) limitations to the generalizability of the resulting model based on the training and testing data that will be used.

A3.3 Define and document Responsible Release Criteria for this Goal. Include:
1) a concise definition of the problem being solved in the intended use,
2) performance metrics and their Responsible Release Criteria, and
3) error types and their Responsible Release Criteria.

A3.4 Document an evaluation plan for each of the performance metrics and error types.
Tags: Ongoing Evaluation Checkpoint.

A3.5 Use the methods defined in requirement A3.4 to conduct evaluations. Document the pre-release results of the evaluations. Determine and document how often ongoing evaluation should be conducted to continue supporting this Goal.
Tags: Ongoing Evaluation Checkpoint.

A3.6 Provide documentation to customers which describes the system's:
1) intended uses, and
2) evidence that the system is fit for purpose for each intended use.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.
Tags: Transparency Note.

A3.7 If an intended use is not supported by evidence, or if evidence comes to light that refutes that the system is fit for purpose for the intended use at any point in the system's use:
1) remove the intended use from customer-facing materials and make current customers aware of the issue, take action to close the identified gap, or discontinue the system,
2) revise documentation related to the intended use, and
3) publish the revised documentation to customers.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

A3.8 Communicate with care about system benefits; follow any applicable guidance from your attorney.

Goal A4: Data governance and management

Microsoft AI systems are subject to appropriate data governance and management practices.

Applies to: All AI systems.

Requirements
A4.1 Define and document data requirements with respect to the system's intended uses, stakeholders, and the geographic areas where the system will be deployed. Document these requirements in the Impact Assessment. **Tags:** Impact Assessment.
A4.2 Define and document procedures for the collection and processing of data, to include annotation, labelling, cleaning, enrichment, and aggregation, where relevant.
A4.3 If you plan to use existing data sets to train the system, assess the quantity and suitability of available data sets that will be needed by the system in relation to the data requirements defined in A4.1. Document this assessment in the Impact Assessment. **Tags:** Impact Assessment.
A4.4 Define and document methods for evaluating data to be used by the system against the requirements defined in A4.1.
A4.5 Evaluate all data sets using the methods defined in requirement A4.4. Document the results of the evaluation.

Goal A5: Human oversight and control

Microsoft AI systems include capabilities that support informed human oversight and control.

Applies to: All AI systems.

Requirements

A5.1 Identify the stakeholders who are responsible for troubleshooting, managing, operating, overseeing, and controlling the system during and after deployment. Document these stakeholders and their oversight and control responsibilities using the Impact Assessment template.
Tags: Impact Assessment.

A5.2 Identify the system elements (including system UX, features, alerting and reporting functions, and educational materials) necessary for stakeholders identified in requirement A5.1 to effectively understand their oversight responsibilities and carry them out. Stakeholders must be able to understand:

1) the system's intended uses,
2) how to effectively execute interactions with the system,
3) how to interpret system behavior,
4) when and how to override, intervene, or interrupt the system, and
5) how to remain aware of the possible tendency of over-relying on outputs produced by the system ("automation bias").

Document the system design elements that will support relevant stakeholders for each oversight and control function.

A5.3 When possible, design the system elements identified in A5.2. When this is not possible (for example, when Microsoft is not responsible for the system UX), provide guidance on human oversight considerations to the third party responsible for implementing the system elements identified in A5.2.

A5.4 Define and document the method to be used to evaluate whether each oversight or control function can be accomplished by stakeholders in realistic conditions of system use. Include the metrics or rubrics that will be used in the evaluations. When this is not possible (for example, when Microsoft is not responsible for oversight and control functions), provide guidance on evaluating oversight and control functions to the third party responsible for evaluating oversight or control functions.

A5.5 Define and document Responsible Release Criteria to achieve this Goal.

A5.6 Conduct evaluations defined by requirement A5.4 using a near-release version of the system. Document the results.

A5.7 If there are Responsible Release Criteria for metrics or rubrics that have not been met, consult with the reviewers named in the Impact Assessment, and in the case of Sensitive Uses, with the Office of Responsible AI, to develop a plan detailing how the gap will be managed until it can be closed. Document that plan.

Tools and practices

Recommendation A5.3.1 Follow the Guidelines for Human-AI Interaction when designing the system.

Recommendation A5.4.1 Assign user researchers to design these evaluations.

Transparency Goals

Goal T1: System intelligibility for decision making

Microsoft AI systems that inform decision making by or about people are designed to support stakeholder needs for intelligibility of system behavior.

Applies to: All AI systems when the intended use of the generated outputs is to inform decision making by or about people.

Requirements
T1.1 Identify: 1) stakeholders who will use the outputs of the system to make decisions, and 2) stakeholders who are subject to decisions informed by the system. Document these stakeholders using the Impact Assessment template. **Tags:** Impact Assessment.
T1.2 Design the system, including, when possible, the system UX, features, reporting functions, and educational materials, so that stakeholders identified in requirement T1.1 can: 1) understand the system's intended uses, 2) interpret relevant system behavior effectively (i.e., in a way that supports informed decision making), and 3) remain aware of the possible tendency of over-relying on outputs produced by the system ("automation bias"). For the two categories of stakeholders identified in requirement T1.1, document: 1) how the system design will support their understanding of the system's intended uses, and 2) how the system aids their ability to interpret relevant system responses, and 3) how the system design discourages automation bias.
T1.3 Define and document the method to be used to evaluate whether each stakeholder who will make decisions or be subject to decisions based on the behavior of the system can interpret the relevant system responses reasonably well. Include the metrics or rubrics that will be used in the evaluations. **Tags:** Ongoing Evaluation Checkpoint.
T1.4 Define and document a Responsible Release Plan, to include Responsible Release Criteria to achieve this Goal. **Tags:** Ongoing Evaluation Checkpoint.
T1.5 Conduct evaluations defined by requirement T1.3. Document the pre-release results of the evaluations. Determine and document how often ongoing evaluation should be conducted to continue supporting this Goal. **Tags:** Ongoing Evaluation Checkpoint.
T1.6 If there are Responsible Release Criteria for metrics or rubrics that that have not been met, consult with the reviewers named in the Impact Assessment, and in the case of Sensitive Uses, with the Office of Responsible AI, to develop a plan detailing how the gap will be managed until it can be closed. Document that plan.

Tools and practices
Recommendation T1.2.1 Follow the Guidelines for Human-AI Interaction when designing the system. **Recommendation T1.2.2** Use one or more techniques available as part of the Interpret ML toolkit to understand the impact of features on system behavior. This may help stakeholders who need to understand model predictions.
Recommendation T1.3.1 Assign user researchers to define, design, and prioritize evaluations in appropriately realistic contexts of use.

Goal T2: Communication to stakeholders

Microsoft provides information about the capabilities and limitations of our AI systems to support stakeholders in making informed choices about those systems.

Applies to: All AI systems.

Requirements

T2.1 Identify:
1) stakeholders who make decisions about whether to employ a system for particular tasks, and
2) stakeholders who develop or deploy systems that integrate with this system.

Document these stakeholders in the Impact Assessment template.
Tags: Impact Assessment.

T2.2 Publish documentation for the system so that stakeholders defined in T2.1 can understand the system. Include:
1) capabilities,
2) intended uses,
3) uses that require extra care or guidance,
4) operational factors and settings that allow for effective and responsible system use,
5) limitations, including uses for which the system was not designed or evaluated, and
6) evidence of system accuracy and performance as well as a description of the extent to which these results are generalizable across use cases that were not part of the evaluation.

When the system is a platform service made available to external customers or partners, a Transparency Note is required.
Tags: Transparency Note.

T2.3 Review and update documentation annually or when any of the following events occur:
1) new uses are added,
2) functionality changes,
3) the product moves to a new release stage,
4) new information about reliable and safe performance becomes known as defined by requirement RS3.3, or
5) new information about system accuracy and performance becomes available.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.
Tags: Transparency Note.

Goal T3: Disclosure of AI interaction

Microsoft AI systems are designed to inform people that they are interacting with an AI system or are using a system that generates or manipulates image, audio, or video content that could falsely appear to be authentic.

Applies to: AI systems that impersonate interactions with humans, unless it is obvious from the circumstances or context of use that an AI system is in use. AI systems that generate or manipulate image, audio, or video content that could falsely appear to be authentic.

Requirements
T3.1 Identify stakeholders who will use or be exposed to the system, in accordance with the Impact Assessment requirements. Document these stakeholders using the Impact Assessment template. **Tags:** Impact Assessment.
T3.2 Design the system, including system UX, features, reporting functions, educational materials, and outputs so that stakeholders identified in T3.1 will be informed of the type of AI system they are interacting with or exposed to. Ensure that any image, audio, or video outputs that are intended to be used outside the system are labelled as being produced by AI.
T3.3 Define and document the method to be used to evaluate whether each stakeholder identified in T3.1 is informed of the type of AI system they are interacting with or exposed to. **Tags:** Ongoing Evaluation Checkpoint.
T3.4 Define and document Responsible Release Criteria to achieve this Goal. **Tags:** Ongoing Evaluation Checkpoint.
T3.5 Conduct evaluations defined by requirement T3.3. Document the pre-release results of the evaluations. Determine and document how often ongoing evaluation should be conducted to continue supporting this goal. **Tags:** Ongoing Evaluation Checkpoint.

Fairness Goals

Goal F1: Quality of service

Microsoft AI systems are designed to provide a similar quality of service for identified demographic groups, including marginalized groups.

Applies to: AI systems when system users or people impacted by the system with different demographic characteristics might experience differences in quality of service that Microsoft can remedy by building the system differently.

Requirements

F1.1 Identify and prioritize demographic groups, including marginalized groups, that may be at risk of experiencing worse quality of service based on intended uses and geographic areas where the system will be deployed. Include:

1) groups defined by a single factor, and
2) groups defined by a combination of factors.

Document the prioritized identified demographic groups using the Impact Assessment template.
Tags: Impact Assessment.

F1.2 Evaluate all data sets to assess inclusiveness of identified demographic groups and collect data to close gaps. Document this process and its results.

F1.3 Define and document the evaluation that you will perform to support this Goal. Include:

1) any system components to be evaluated, in addition to the whole system,
2) the metrics to be used to evaluate the system components and the whole system, and
3) a description of the data set to be used for this evaluation.

Tags: Ongoing Evaluation Checkpoint.

F1.4 Define and document Responsible Release Criteria to achieve this Goal, as follows:
For each metric, document:

1) any target minimum performance level for all groups, and
2) the target maximum (absolute or relative) performance difference between groups.

Tags: Ongoing Evaluation Checkpoint.

F1.5 Evaluate the system according to the defined Responsible Release Criteria.
Tags: Ongoing Evaluation Checkpoint.

F1.6 Reassess the system design, including the choice of training data, features, objective function, and training algorithm, to pursue the goals of:

1) improving performance for any identified demographic group that does not meet any target minimum performance level, and
2) minimizing performance differences between identified demographic groups, paying particular attention to those that exceed the target maximum, while recognizing that doing so may appear to affect system performance and that it is seldom clear how to make such tradeoffs.

Consult with your attorney to determine your approach to this, including how you will identify and document tradeoffs.
Tags: Ongoing Evaluation Checkpoint.

F1.7 Identify and document any justifiable factors, such as circumstantial and other operational factors (e.g., "background noise" for speech recognition systems or "image resolution" for facial recognition systems), that account for:

1) any inability to meet any target minimum performance level for any identified demographic group, and
2) any remaining performance differences between identified demographic groups.

Tags: Ongoing Evaluation Checkpoint.

F1.8 Document the pre-release results from requirements F1.4, F1.5, and F1.6. Determine and document how often ongoing evaluation should be conducted to continue supporting this Goal.
Tags: Ongoing Evaluation Checkpoint.

F1.9 Publish information for customers about:

1) identified demographic groups for which performance may not meet any target minimum performance level,
2) any remaining performance disparities between identified demographic groups that may exceed the target maximum, and
3) any justifiable factors that account for these performance levels and differences.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.
Tags: Transparency Note.

Tools and practices

Recommendation F1.1.1 For identifying people by age, gender identity, and ancestry in North America, use Best Practices for Age, Gender Identity, and Ancestry.

Recommendation F1.1.2 Work with user researchers to understand variations in demographic groups across intended uses and geographic areas.

Recommendation F1.1.3 Work with domain-specific subject matter experts to understand the factors that impact performance of your system and how they vary across identified demographic groups in this domain.

Recommendation F1.1.4 Work with members of identified demographic groups to understand the risks of and impacts associated with differences in quality of service. Consider using the Community Jury technique to conduct these discussions.

Recommendation F1.2.1 Use Analysis Platform to understand the representation of identified demographic groups in data sets that you plan to use for training and evaluating your system, respecting privacy controls for working with sensitive data.

Recommendation F1.2.2 Document the representation of identified demographic groups in a Datasheet.

Recommendation F1.5.1 Use the Fairlearn Python toolkit's assessment and mitigation capabilities, if appropriate for the system.

Recommendation F1.5.2 Use Error Analysis to help understand factors that may account for performance levels and differences, if appropriate for the system.

Recommendation F1.5.3 Use one or more techniques available as part of the Interpret ML toolkit to help understand factors that may account for performance levels and differences, if appropriate for the system.

Recommendation F1.6.1 Use the Fairlearn Python toolkit's assessment and mitigation capabilities, if appropriate for the system.

Recommendation F1.6.2 Be prepared to collect additional training data for identified demographic groups.

Recommendation F1.7.1 Use Error Analysis to help understand factors that may account for performance levels and differences, if appropriate for the system.

Recommendation F1.7.2 Use one or more techniques available as part of the Interpret ML toolkit to help understand factors that may account for performance levels and differences, if appropriate for the system.

Goal F2: Allocation of resources and opportunities

Microsoft AI systems that allocate resources or opportunities in essential domains are designed to do so in a manner that minimizes disparities in outcomes for identified demographic groups, including marginalized groups.

Applies to: AI systems that generate outputs that directly affect the allocation of resources or opportunities relating to finance, education, employment, healthcare, housing, insurance, or social welfare.

Requirements

F2.1 Identify and prioritize demographic groups, including marginalized groups, that may be at risk of being differentially impacted by the system based on intended uses and geographic areas where the system will be deployed. Include:

1) groups defined by a single factor, and
2) groups defined by a combination of factors.

Document the prioritized identified demographic groups using the Impact Assessment template.
Tags: Impact Assessment.

F2.2 Evaluate all data sets to assess inclusiveness of identified demographic groups and collect data to close any gaps. Document this process and its results.

F2.3 Define and document the evaluation that you will perform to support this Goal. Include:

1) any system components to be evaluated, in addition to the whole system,
2) the metrics to be used to evaluate the system components and the whole system, and
3) the data set to be used for this evaluation.

Tags: Ongoing Evaluation Checkpoint.

F2.4 Define and document Responsible Release Criteria to achieve this Goal, as follows:
For each metric, document the target maximum difference (absolute or relative) between the rates at which resources and opportunities are allocated to groups.
Tags: Ongoing Evaluation Checkpoint.

F2.5 Evaluate the system according to the defined Responsible Release Criteria.
Tags: Ongoing Evaluation Checkpoint.

F2.6 Reassess the system design, including the choice of training data, features, objective function, and training algorithm, to pursue the goal of minimizing differences between the rates at which resources and opportunities are allocated to identified demographic groups, paying particular attention to those that exceed the target maximum difference, while recognizing that doing so may appear to affect system performance and it is seldom clear how to make such trade-offs.
Consult with your attorney to determine your approach to this, including how you will identify and document trade-offs.
Tags: Ongoing Evaluation Checkpoint.

F2.7 Identify and document any justifiable factors that account for any remaining differences between the rates at which resources and opportunities are allocated to identified demographic groups.
Tags: Ongoing Evaluation Checkpoint.

F2.8 Document the pre-release results for the evaluation described by requirements F2.4, F2.5, and F2.6. Determine and document how often ongoing evaluation should be conducted to continue supporting this goal.
Tags: Ongoing Evaluation Checkpoint.

F2.9 Publish information for customers about:

1) any remaining differences between the rates at which resources and opportunities are allocated to identified demographic groups, and
2) any justifiable factors that account for these differences. When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.

Tags: Transparency Note.

Tools and practices

Recommendation F2.1.1 For North America, use Best Practices for Age, Gender Identity, and Ancestry to help identify demographic groups and methods for collecting demographic information.

Recommendation F2.1.2 Work with user researchers to understand variations in demographic groups across intended uses and geographic areas.

Recommendation F2.1.3 Work with domain-specific subject matter experts to understand the facts that impact performance of your system and how they vary across identified demographic groups in this domain.

Recommendation F2.1.4 Work with members of identified demographic groups to understand risks of and impacts associated with differences between the rates at which resources and opportunities are allocated.

Recommendation F2.2.1 Use Analysis Platform to understand the representation of identified demographic groups, respecting privacy requirements for using sensitive data.

Recommendation F2.2.2 Document the representation of identified demographic groups in a Datasheet.

Recommendation F2.5.1 Use the Fairlearn Python toolkit's assessment and mitigation capabilities, if appropriate for the system.

Recommendation F2.5.2 Use Error Analysis to help understand factors that may account for differences between the rates at which resources and opportunities are allocated to the identified demographic groups, if appropriate for the system.

Recommendation F2.5.3 Use one or more techniques available as part of the Interpret ML toolkit to help understand factors that may account for differences between the rates at which resources and opportunities are allocated to the identified demographic groups, if appropriate for the system.

Recommendation F2.6.1 Use the Fairlearn Python toolkit's assessment and mitigation capabilities, if appropriate for the system.

Recommendation F2.7.1 Use Error Analysis to help understand factors that may account for differences between the rates at which resources and opportunities are allocated to the identified demographic groups, if appropriate for the system.

Recommendation F2.7.2 Use Interpret ML to help understand factors that may account for differences between the rates at which resources and opportunities are allocated to the identified demographic groups, if appropriate for the system.

Goal F3: Minimization of stereotyping, demeaning, and erasing outputs

Microsoft AI systems that describe, depict, or otherwise represent people, cultures, or society are designed to minimize the potential for stereotyping, demeaning, or erasing identified demographic groups, including marginalized groups.

Applies to: AI systems when system outputs include descriptions, depictions, or other representations of people, cultures, or society.

Requirements

F3.1 Identify and prioritize demographic groups, including marginalized groups, that may be at risk of being subject to stereotyping, demeaning, or erasing outputs of the system. Include:

1) groups defined by a single factor, and
2) groups defined by a combination of factors.

Document the prioritized identified demographic groups using the Impact Assessment template.
Tags: Impact Assessment.

F3.2 Define and document any system components to be evaluated, in addition to the whole system.

F3.3 Define and document a plan to evaluate the system components and the whole system for risks of stereotyping, demeaning, and erasing the prioritized identified demographic groups.
Tags: Ongoing Evaluation Checkpoint.

F3.4 Evaluate the system according to the plan defined in requirement F3.3.
Tags: Ongoing Evaluation Checkpoint.

F3.5 Reassess the system design, including the choice of training data, features, objective function, and training algorithm, to pursue the goal of minimizing the potential for stereotyping, demeaning, and erasing the identified demographic groups.
Tags: Ongoing Evaluation Checkpoint.

F3.6 Document the pre-release results from requirements F3.4 and F3.5. Determine and document how often ongoing evaluation should be conducted to continue supporting this goal.
Tags: Ongoing Evaluation Checkpoint.

F3.7 Publish information for customers about these risks involving identified demographic groups. When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.
Tags: Transparency Note.

Tools and practices

Recommendation F3.1.1 Work with user researchers, subject matter experts, and members of identified demographic groups to understand these risks and their impacts.

Recommendation F3.4.1 Use CheckList to help evaluate these risks involving identified demographic groups, if appropriate for the system.
Recommendation F3.4.2 Use red teaming exercises to evaluate these risks involving identified demographic groups.

Recommendation F3.5.1 Mitigate any risks of these types of harms that you can. In addition, establish feedback mechanisms and a plan for addressing problems, in alignment with Reliability and Safety Goal RS3. Note that this approach is recommended in acknowledgment of the fact that the state-of-the-art in mitigating these risks is less advanced than the state-of-the-art in mitigating differences in quality of service or allocative harms.

Reliability & Safety Goals

Goal RS1: Reliability and safety guidance

Microsoft evaluates the operational factors and ranges within which AI systems are expected to perform reliably and safely, remediates issues, and provides related information to customers.

Applies to: All AI systems.

Requirements
RS1.1 Document how: 1) reliable and safe behavior is defined for this system and, 2) what acceptable error rates are for overall system performance in the context of intended uses. **Tags:** Ongoing Evaluation Checkpoint.
RS1.2 Evaluate training and test data sets to ensure that they include representation of the intended uses, operational factors, and an appropriate range of settings for each factor. Document the evaluation. **Tags:** Ongoing Evaluation Checkpoint.
RS1.3 Determine and document the operational factors, including quality of system input, use, and operational context that are critical to manage for reliable and safe use of the system in its deployed context. **Tags:** Ongoing Evaluation Checkpoint.
RS1.4 Define and document acceptable ranges for each operational factor important to support reliable and safe system use. Define and document an acceptable error rate for the system when operating within these ranges. **Tags:** Ongoing Evaluation Checkpoint.
RS1.5 Define intended uses, if any, where additional operational factors, more narrow or different acceptable ranges, or lower acceptable error rates (including false positive and false negative error rates), are advised to ensure reliability and safety. Document your conclusions. **Tags:** Ongoing Evaluation Checkpoint.
RS1.6 Define and document an evaluation plan based on requirements RS1.1, RS1.3, RS1.4, and RS1.5, to include the environment in which the system will be evaluated. **Tags:** Ongoing Evaluation Checkpoint.
RS1.7 Evaluate the system according to the evaluation plan defined in requirement RS1.6 to ensure reliable and safe system behavior. Document the pre-release results of the evaluation. Determine and document how often ongoing evaluation should be conducted to continue supporting this goal. **Tags:** Ongoing Evaluation Checkpoint.
RS1.8 In the event of failure cases within operational factors and defined ranges, work to resolve the issues. If the Responsible Release Criteria established in requirements RS1.1, RS1.3, RS1.4, and RS1.5 cannot be met, a reassessment of intended uses and updated documentation is required. **Tags:** Ongoing Evaluation Checkpoint.
RS1.9 Provide documentation to customers and potential customers of the system that includes the outputs of requirements RS1.2, RS1.7 and RS1.8, and any unsupported uses defined in the Impact Assessment and in RS1.8. When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note. **Tags:** Impact Assessment, Transparency Note.

Tools and practices

Recommendation RS1.1.1 Interview safety experts and review relevant literature for domains where the system may impact the safety of people.

Recommendation RS1.4.1 Interview customers to understand operational factors and their variations.

Goal RS2: Failures and remediations

Microsoft AI systems are designed to minimize the time to remediation of predictable or known failures.

Applies to: All AI systems.

Requirements

RS2.1 Define predictable failures, including false positive and false negative results for the system as a whole and how they would impact stakeholders for each intended use. Use the Impact Assessment template to document any adverse impacts of these failures on stakeholders.
Tags: Impact Assessment.

RS2.2 For each case of a predictable failure likely to have an adverse impact on a stakeholder, document the failure management approach:

1) When possible, design and build the system to avoid this failure. Describe the design solution. Estimate the time range for resolving predictable failures for each designed solution or indicate that the failure will be prevented by design.
2) When a failure cannot be prevented by design, build a fallback option that may be used when this failure occurs. Describe the fallback option and document the estimated time required to invoke and use the fallback option.
3) Provide training and documentation for stakeholders accountable for system oversight that supports their resolution of the failure. Describe the documentation and training.

RS2.3 Document your plan for managing previously unknown failures that come to light once the system is in use:

1) Describe the system's rollback plan and document the time that may elapse until the entire system, across all endpoints can be rolled back.
2) Describe support for turning features off and document the time that may elapse until the feature can be turned off across all endpoints.
3) Describe the process for updating and releasing updates to each model and document the time that may elapse until the system has been updated across all endpoints.
4) Describe how customers, partners, and end users will be notified of changes to the system, updated understandings of failures, and their best mitigations.

RS2.4 Provide training and documentation for system owners, developers, customer support and other stakeholders responsible for managing the system to support their remediation and mitigation of predictable failures identified in requirement RS2.1. Document the training and documentation provided.

Tools and practices

Recommendation RS2.1.1 Conduct Failure Mode and Effects Analysis.

Recommendation RS2.2.1 Follow the Guidelines for Human-AI Interaction when designing the system to help manage failures.

Goal RS3: Ongoing monitoring, feedback, and evaluation

Microsoft AI systems are subject to ongoing monitoring, feedback, and evaluation so that we can identify and review new uses, identify and troubleshoot issues, manage and maintain the systems, and improve them over time.

Applies to: All AI systems.

Requirements

RS3.1 Establish and document a detailed inventory of the system health monitoring methods to be used, to include:

1) data and insights generated from data repositories, system analytics, and associated alerts,
2) processes by which customers can submit information about failures and concerns, and
3) processes by which the general public can submit feedback.

RS3.2 Define and document a standard operating procedure and system health monitoring action plan for each monitoring channel for the system, to include:

1) processes for reproducing system failures to support troubleshooting and prevention of future failures,
2) which events will be monitored,
3) how events will be prioritized for review,
4) the expected frequency of those reviews,
5) how events will be prioritized for response and timing to resolution,
6) how high priority issues related to supporting the Standard and its goals will be escalated to the Office of Responsible AI, and
7) engaging customer service to ensure that they are aware of how to respond to issues for the system.

RS3.3 When new uses, critical operational factors, or changes in the supported range of an operational factor are identified, determine whether any new use or operational factor can be supported with the existing system, will be supported but require additional work, or will not be supported.

- When new uses or operational factors identified are to be supported, evaluate the updated system in accordance with requirement RS1.6, add the new intended use to the Impact Assessment, and publish updated communication in accordance with requirement RS1.9.
- When these new uses or operational factor range changes cannot or will not be accommodated to ensure reliable and safe performance of the system update customer documentation described in RS1.9 to include the new use as an unsupported use.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.
Tags: Impact Assessment, Transparency Note.

RS3.4 When a system is to be used for a Sensitive Use that imposes qualification or quality control requirements beyond the intended uses and/or operational factor ranges, conduct an evaluation specific to this use. If the required Responsible Release Criteria cannot be met, the Office of Responsible AI will review the results and decide how to proceed. Document any changes to the Responsible Release Criteria and document the results of evaluation.

RS3.5 Conduct all evaluations tagged as Ongoing Evaluation Checkpoints in other Goals on an ongoing basis.

RS3.6 If there are targets in Ongoing Evaluation Checkpoints that are no longer satisfied, consult with named reviewers, and in the case of Sensitive Uses, with the Office of Responsible AI, to develop and implement a plan to close any gaps. Document the process, its results, and conclusions.

RS3.7 If evidence comes to light that refutes the system is fit for purpose for an intended use at any point in the system's use:

1) remove the intended use from customer-facing materials and make current customers aware of the issue, take action to close the identified gap, or discontinue the system,
2) revise documentation related to the intended use, and
3) publish the revised documentation to customers.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.
Tags: Transparency Note.

RS3.8: Review and update documentation required by Goal T2 when any of the following events occur:

1) new uses are added,
2) functionality changes,
3) new information about reliable and safe performance becomes known as defined by requirement RS3.3, or
4) new information about system accuracy and performance becomes available.

When the system is a platform service made available to external customers or partners, include this information in the required Transparency Note.
Tags: Transparency Note.

RS3.9: Escalate unresolved issues related to supporting the Standard and its requirements to the Office of Responsible AI.

Privacy & Security Goals

Goal PS1: Privacy Standard compliance

Microsoft AI systems are designed to protect privacy in accordance with the Microsoft Privacy Standard.

Applies when: Microsoft Privacy Standard applies.

Goal PS2: Security Policy compliance

Microsoft AI systems are designed to be secure in accordance with the Microsoft Security Policy.

Applies when: Microsoft Security Policy applies.

Inclusiveness Goal

Goal I1: Accessibility Standards compliance

Microsoft AI systems are designed to be inclusive in accordance with the Microsoft Accessibility Standards.

Applies when: Microsoft Accessibility Standards apply.

Scan this code to access responsible AI resources from Microsoft: