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.