# Document Metadata

author: jfraze@mycomp.org

title: Transparency and Explainability in GenAI Use

nist\_function: Measure

priority\_phase: Must

last\_reviewed: 2025-06-24

status: draft

## **Purpose**

To promote responsible deployment of GenAI by requiring that outputs, system behavior, and underlying decision paths be interpretable by human reviewers, especially in clinical or patient-facing contexts.

## **Scope**

Covers all deployed GenAI systems used for summarization, decision support, triage, documentation, or direct communication in hospital settings.

## **Policy Statement**

All generative AI outputs that affect patient care or hospital operations must be explainable in a manner appropriate to the target audience (e.g., clinicians, patients, regulators). Explainability measures should be technical (model-level) and contextual (use-level).

## **Roles and Responsibilities**

• AI Ethics Committee: Define organizational transparency thresholds and acceptability criteria.

• Developers & Integrators: Embed explanation capabilities, such as rationale snippets or attention maps.

• Clinical Stakeholders: Validate explanations for utility in care workflows and patient trust.

## **Implementation Phases**

### **Must Do**

• Enable user-accessible rationale or justification snippets (AI 600-1 §2.4.1).

• Mark model-generated content with visible indicators (AI RMF MEASURE-2.6).

• Provide documentation on how outputs are produced and ranked.

### **Should Do**

• Use saliency or attention visualizations for internal model interpretability.

• Train staff on interpreting GenAI outputs and known failure modes.

• Integrate explainability with audit trails and risk scoring systems.

### **Recommended**

• Provide explainable layers for regulators and auditors.

• Tailor explainability modes to patient literacy and accessibility needs.