Quick Reference Guide: SHAP Values

■ What Are SHAP Values?

SHAP = SHapley Additive exPlanations. They explain how each input feature contributes to an AI model’s prediction. Think of it like splitting a bill fairly: each feature pays its share toward the final decision.

■ Key Principles

• Baseline Prediction = Model’s average prediction with no feature info.
• SHAP Value = How much a feature pushes the prediction up (+) or down (–) compared to the baseline.
• Positive SHAP → increases chance of outcome.
• Negative SHAP → decreases chance of outcome.
• SHAP is model-specific and dataset-dependent (values are not universal thresholds).

■ Visual Example (Loan Approval Model)

Baseline Probability: 0.55 (55%)
Final Prediction: 0.80 (80%) → Loan Approved

Baseline (0.55)		
+ Income (+0.25)	↑ Positive	Increased chance of approval
+ Stable Employment (+0.10)	↑ Positive	Supportive factor
– High Debt Ratio (–0.05)	↓ Negative	Decreased approval likelihood
– Credit Score (–0.05)	↓ Negative	Weakened approval
Final Prediction = 0.80 (80%)		

■ Why It Matters for Governance

• Ensures fairness (detect bias, explain rejections).
• Provides audit trail (why a decision was made).
• Meets regulatory requirements (MAS FEAT, FCA, EU AI Act).
• Builds trust with customers (clear, human-readable reasons).

■ How to Use This Guide

• Use SHAP reports for model validation before deployment.
• Attach SHAP explanations to case reviews (loan denials, fraud alerts).
• Keep SHAP documentation in the Explainability Register for audit/regulatory checks.