### Assurance Case Goal: Interpretability of the DeepMind Machine Learning Enabled System

- \*\*G1: Demonstrate that the DeepMind Machine Learning (ML) system is interpretable\*\*

- \*\*S1: Definition of Interpretability\*\*

- G1.1: Define key interpretability metrics applicable to the DeepMind ML system

- G1.2: Identify interpretability requirements specific to the DeepMind ML system's application domain

- \*\*S2: Design Considerations for Interpretability\*\*

- G2.1: Design and implement features that ensure model transparency

- G2.2: Incorporate human-understandable explanations for model decisions

- \*\*S3: Application of Interpretability Techniques\*\*

- G3.1: Use established interpretability frameworks (e.g., LIME, SHAP) across DeepMind ML systems

- G3.2: Validate effectiveness of interpretability methods through user studies

- \*\*S4: Testing for Interpretability\*\*

- G4.1: Conduct case studies to assess the clarity of model explanations

- SN4.1.1: Report on qualitative feedback from test participants

- SN4.1.2: Evaluate the consistency of explanations across different model runs

- G4.2: Quantitative testing of interpretability metrics

- SN4.2.1: Measure fidelity and perturbation stability of explainability techniques

- SN4.2.2: Test robustness of interpretations under model re-training scenarios

- \*\*S5: Documentation and Reporting\*\*

- G5.1: Create comprehensive documentation detailing interpretability features and methods

- G5.2: Report on limitations and potential risks in current interpretability tools and techniques

- \*\*S6: Compliance and Standard Adherence\*\*

- G6.1: Ensure compliance with relevant AI ethics and transparency standards

- G6.2: Evaluate and update interpretability protocols in line with evolving standards

This tree presents the hierarchical goals (G), strategies (S), and sub-goals/sub-nodes (SN) designed to ensure the interpretability of a sophisticated system like DeepMind's ML system. Each element serves to build a comprehensive case for the system's interpretability by addressing both the technical and user-centric aspects of the feature.