**Appendix B**

Paired AUC Comparison and DeLong Test for Each Theoretical Model Within and Across Stages

To assess the statistical significance of differences in model performance, we employ the DeLong test for comparing the areas under the ROC curves (AUCs) of paired classification models. The DeLong test is a non-parametric method commonly used to evaluate whether observed differences in AUCs between two models or classifiers are statistically significant, applied to the same dataset. This appendix reports pairwise comparisons of AUCs for all theoretical models reported in the paper, both within the conflict onset and escalation stages, and across logistic regression and random forest models.

Table 1 below presents the results of the DeLong test comparing the logistic regression and random forest models during the conflict onset stage. The differences in AUC between the models are all statistically significant.

**Table 1: DeLong Test Results Comparing Logistic Regression and Random Forest Models for Conflict Onset**

**A screenshot of a graph

AI-generated content may be incorrect.**

Table 2 reports the DeLong test results for all pairwise comparisons of AUC scores from the Random Forest models predicting conflict onset. With the exception of comparisons involving the Resource Mobilization and Political Opportunity + Resource Mobilization models, as well as two instances involving the Complete Model, the differences in AUCs are not statistically significant.

**Table 2: DeLong Test Results for Pairwise Comparisons of Random Forest Model AUCs During the Conflict Onset Stage**

**A table of numbers and letters

AI-generated content may be incorrect.**

Table 3 presents the DeLong test results comparing logistic regression and random forest models during the conflict escalation stage. The differences in AUCs between the models are statistically significant in nearly all cases.

**Table 3: DeLong Test Results Comparing Logistic Regression and Random Forest Models for Conflict Escalation**

**A screenshot of a graph

AI-generated content may be incorrect.**

Table 4 reports the DeLong test results for all pairwise comparisons of AUC scores from the Random Forest models predicting conflict escalation. The differences in AUCs are not statistically significant.

**Table 4: DeLong Test Results for Pairwise Comparisons of Random Forest Model AUCs in the Conflict Escalation Stage**

**A screenshot of a computer screen

AI-generated content may be incorrect.**