### Introduction

In many military operations, we must be able to explain how an A.I. algorithm reached a decision. Previously, we argued that **the utility of an explanation depends on** the **nature of the task** being performed **and the** **role of the agent** consuming the explanation [1].

To explore this, **we have developed a framework of datasets, machine learning models and explanation techniques** which allows for the comparison between a range of explanations in the context of different tasks.

Objectives of the Framework:

**Build intuition for current and future explanation techniques** allowing for innovations in their use and in the creation of novel techniques.

**Produce data that can be used to develop metrics** to measure utility of explanation techniques and benchmark their resource cost.

Figure 1: Using the interpretability framework to compare different explanation techniques generated relating to the same classification of an input image.

**Experiment with ‘coalitions’ of machine learning & explanation services** in a dynamic context to satisfy mission requirements with consideration of resource usage.

Explanation comparison placeholder

#### Intended Uses of the Framework

**Comparative/sensitivity analyses of explanation techniques** e.g., using the Framework, we demonstrated the debugging process of generating multiple explanations from the same technique (LIME) to measure and improve its stability [2].

**Empirical studies of explanation visualisations** Existing explanation techniques are subject to post-hoc interpretation by the recipient. This can be greatly affected by the visualization and presentation of the explanation and can lead to the recipient projecting their existing assumptions on to the explanation. Future work will explore this issue.

**Eliciting stakeholder requirements for application-level interpretability** We will use the framework to engage with subject-matter experts to gain insights as to what constitutes a useful explanation for aiding task performance.

#### References

#### [1] Tomsett et al. “Interpretable to whom? A role-based model for analyzing interpretable machine learning systems” in 3rd Annual Workshop on Human Interpretability in Machine Learning (WHI 2018)

#### [2] Stiffler et al. “An Analysis of Reliability Using LIME with Deep Learning Models” in DAIS ITA AFM 2018